



Checkpointing strategies with prediction windows

Guillaume Aupy, Yves Robert, Frédéric Vivien, Dounia Zaidouni

**RESEARCH
REPORT**

N° 8239

February 2013

Project-Team ROMA



Checkpointing strategies with prediction windows

Guillaume Aupy*, Yves Robert^{*†‡}, Frédéric Vivien^{§*}, Dounia Zaidouni^{§*}

Project-Team ROMA

Research Report n° 8239 — February 2013 — 41 pages

Abstract: This paper deals with the impact of fault prediction techniques on checkpointing strategies. We suppose that the fault-prediction system provides prediction windows instead of exact predictions, which dramatically complicates the analysis of the checkpointing strategies. We propose a new approach based upon two periodic modes, a regular mode outside prediction windows, and a proactive mode inside prediction windows, whenever the size of these windows is large enough. We are able to compute the best period for any size of the prediction windows, thereby deriving the scheduling strategy that minimizes platform waste. In addition, the results of this analytical evaluation are nicely corroborated by a comprehensive set of simulations, which demonstrate the validity of the model and the accuracy of the approach.

Key-words: Fault-tolerance, checkpointing, prediction, algorithms, model, exascale

* LIP, École Normale Supérieure de Lyon, France

† University of Tennessee Knoxville, USA

‡ Institut Universitaire de France

§ INRIA

**RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES**

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Protocoles de checkpoint et intervalles de prédiction de fautes

Résumé : Ce travail considère l'impact des techniques de prédiction de fautes sur les stratégies de protocoles de sauvegarde de points de reprise (*check-points*) et de redémarrage. Nous supposons que le système de prédiction de faute fournit, en plus de ses caractéristiques, rappel (taux de pannes prévues sur nombre total de pannes) et précision (taux de vraies pannes parmi le nombre total de pannes annoncées), des intervalles dans lesquels les fautes risquent d'arriver plutôt qu'une date exacte, ce qui complique énormément l'analyse des stratégies de checkpoint. Nous proposons une nouvelle approche qui se base sur deux modes périodiques, le mode régulier (en dehors des fenêtres de prédiction), et le mode proactif (à l'intérieur des fenêtres de prédiction), quand ces fenêtres sont suffisamment larges. Nous sommes capables de calculer la période optimale pour n'importe quelle taille de fenêtre de prédiction, dérivant ainsi la stratégie d'ordonnancement optimale pour minimiser les pertes. Enfin, ces résultats théoriques sont corroborés par un ensemble complet de simulations, ce qui démontre la validité du modèle et la précision de notre approche.

Mots-clés : Tolérance aux pannes, checkpoint, prédiction, algorithmes, modèle, exascale

1 Introduction

In this paper, we assess the impact of fault prediction techniques on checkpointing strategies. We assume to have jobs executing on a platform subject to faults, and we let μ be the mean time between faults (MTBF) of the platform. In the absence of fault prediction, the standard approach is to take periodic checkpoints, each of length C , every period of duration T . In steady-state utilization of the platform, the value T_{opt} of T that minimizes the expected waste of resource usage due to checkpointing is easily approximated as $T_{\text{opt}} = \sqrt{2\mu C} + C$, or $T_{\text{opt}} = \sqrt{2(\mu + R)C} + C$ (where R is the duration of the recovery). The former expression is the well-known Young's formula [1], while the latter is due to Daly [2].

Assume now that some fault prediction system is available. Such a system is characterized by two critical parameters, its recall r , which is the fraction of faults that are indeed predicted, and its precision p , which is the fraction of predictions that are correct (i.e., correspond to actual faults). In the simple case where predictions are exact-date predictions, several recent papers [3, 4] have independently shown that the optimal checkpointing period becomes $T_{\text{opt}} = \sqrt{\frac{2\mu C}{1-r}}$. This latter expression is valid only when μ is large enough and can be seen as an extension of Young's formula where μ is replaced by $\frac{\mu}{1-r}$: faults are replaced by non-predicted faults, and the overhead due to false predictions is negligible. A more accurate expression for the optimal checkpointing period is available in [4].

This paper deals with the realistic case (see [5, 6] and Section 5) where the predictor system does not provide exact dates for predicted events, but instead provides *prediction windows*. A *prediction window* is a time interval of length I during which the predicted event is likely to happen. Intuitively, one is more at risk during such an interval than in the absence of any prediction, hence the need to checkpoint more frequently. But with which period? And what is the size of the prediction window above which it proves worthwhile to use a different (smaller) checkpointing period?

The main objective of this paper is to provide a quantitative answer to these questions. Our key contributions are the following: (i) The design of several checkpointing policies that account for the different sizes of prediction windows; (ii) The analytical characterization of the best policy for each set of parameters; and (iii) The validation of the theoretical results via extensive simulations, for both Exponential and Weibull failure distributions. It turns out that the analysis of the waste is dramatically more complicated than when using exact-date predictions [3, 4].

The rest of the paper is organized as follows. First we detail the framework in Section 2. In Section 3 we describe the new checkpointing policies with prediction windows, and show how to compute the optimal checkpointing periods that minimize the platform waste. Section 4 is devoted to simulations. Section 5 provides a brief overview of related work. Finally, we present concluding remarks in Section 6.

2 Framework

2.1 Checkpointing strategy

We consider a *platform* subject to faults. Our work is agnostic of the granularity of the platform, which may consist either of a single processor, or of several processors that work concurrently and use coordinated checkpointing. *Checkpoints* are taken at regular intervals, or periods, of length T . We denote by C the duration of a checkpoint; by construction, we must enforce that $C \leq T$. Useful work is done only during $T - C$ units of time for every period of length T , if no fault occurs. Hence the *waste* due to checkpointing in a fault-free execution is $\text{WASTE} = \frac{C}{T}$. In the following, the *waste* always denote the fraction of time that the platform is not doing useful work.

When a fault strikes the platform, the application is lacking some resource for a certain period of time of length D , the *downtime*. The downtime accounts for software rejuvenation (i.e., rebooting [7, 8]) or for the replacement of the failed hardware component by a spare one. Then, the application recovers from the last checkpoint. R denotes the duration of this *recovery* time.

2.2 Fault predictor

A fault predictor is a mechanism that is able to predict that some faults will take place, within some time-interval window. In this paper, we assume that the predictor is able to generate its predictions early enough so that a *proactive* checkpoint can indeed be taken before or during the event. A first proactive checkpoint will typically be taken just before the beginning of the prediction window, and possibly several other ones will be taken inside the prediction window, if its size I is large enough.

Proactive checkpoints may have a different length C_p than regular checkpoints of length C . In fact there are many scenarios. On the one hand, we may well have $C_p > C$ in scenarios where regular checkpoints are taken at time-steps where the application memory footprint is minimal [9]; on the contrary, proactive checkpoints are taken according to predictions that can take place at arbitrary instants. On the other hand, we may have $C_p < C$ in other scenarios [10], e.g., when the prediction is localized to a particular resource subset, hence allowing for a smaller volume of checkpointed data. To keep full generality, we deal with two checkpoint sizes in this paper: C for periodic checkpoints, and C_p for proactive checkpoints (those taken upon predictions).

The accuracy of the fault predictor is characterized by two quantities, the *recall* and the *precision*. The recall r is the fraction of faults that are predicted while the precision p is the fraction of fault predictions that are correct. Traditionally, one defines three types of *events*: (i) *True positive* events are faults that the predictor has been able to predict (let True_P be their number); (ii) *False positive* events are fault predictions that did not materialize as actual faults (let False_P be their number); and (iii) *False negative* events are faults that were not predicted (let False_N be their number). With these definitions, we have $r = \frac{\text{True}_P}{\text{True}_P + \text{False}_N}$ and $p = \frac{\text{True}_P}{\text{True}_P + \text{False}_P}$.

In the literature, the *lead time* is the interval between the date at which the prediction is made available, and the predicted date of failure (or, more precisely, the beginning of the prediction window). However, because we do

not consider pro-active actions with different durations (they all have length C_p), we point out that the distribution of these lead times is irrelevant to our problem. Indeed, either we have the time to take a proactive action before the failure strikes or not. Therefore, if a failure strikes less than C_p seconds after the prediction is made available, the prediction was useless. In other words, predicted failures that come too early to enable any proactive action should be classified as unpredicted faults, leading to a smaller value of the predictor recall and to a shorten prediction window. Therefore, in the following, we consider, without loss of generality, that all predictions are made available C_p seconds before the beginning of the prediction window.

2.3 Fault rates

The key parameter is μ , the mean time between faults (MTBF) of the platform. If the platform is made of N components whose individual MTBF is μ_{ind} , then $\mu = \frac{\mu_{\text{ind}}}{N}$. This result is true regardless of the fault distribution law[4]. In addition to μ , the platform MTBF, let μ_P be the mean time between predicted events (both true positive and false positive), and let μ_{NP} be the mean time between unpredicted faults (false negative). Finally, we define the mean time between events as μ_e (including all three event types). The relationships between μ , μ_P , μ_{NP} , and μ_e are the following:

- Rate of unpredicted faults: $\frac{1}{\mu_{NP}} = \frac{1-r}{\mu}$, since $1-r$ is the fraction of faults that are unpredicted;
- Rate of predicted faults: $\frac{r}{\mu} = \frac{p}{\mu_P}$, since r is the fraction of faults that are predicted, and p is the fraction of fault predictions that are correct;
- Rate of events: $\frac{1}{\mu_e} = \frac{1}{\mu_P} + \frac{1}{\mu_{NP}}$, since events are either predictions (true or false), or unpredicted faults.

3 Checkpointing strategies

In this section, we introduce the new checkpointing strategies, and we determine the waste that they induce. We then proceed to computing the optimal period for each strategy.

3.1 Description of the different strategies

We consider the following general scheme:

1. While no fault prediction is available, checkpoints are taken periodically with period T ;
2. When a fault is predicted, we decide whether to take the prediction into account or not. This decision is randomly taken: with probability q , we trust the predictor and take the prediction into account, and, with probability $1-q$, we ignore the prediction;
3. If we decide to trust the predictor, we use various strategies, depending upon the length I of the prediction window.

Before describing the different strategies in the situation (3), we point out that the rationale for not always trusting the predictor is to avoid taking useless

checkpoints too frequently. Intuitively, the precision p of the predictor must be above a given threshold for its usage to be worthwhile. In other words, if we decide to checkpoint just before a predicted event, either we will save time by avoiding a costly re-execution if the event does correspond to an actual fault, or we will lose time by unduly performing an extra checkpoint. We need a larger proportion of the former cases, i.e., a good precision, for the predictor to be really useful.

Now, to describe the strategies used when we trust a prediction (situation (3)), we define two *modes* for the scheduling algorithm:

Regular: This is the mode used when no fault prediction is available, or when a prediction is available but we decide to ignore it (with probability $1 - q$). In regular mode, we use periodic checkpointing with period T_R . Intuitively, T_R corresponds to the checkpointing period T of Section 2.1.

Proactive: This is the mode used when a fault prediction is available and we decide to trust it, a decision taken with probability q . Consider such a trusted prediction made with the prediction window $[t_0, t_0 + I]$. Several strategies can be envisioned:

(1) **INSTANT**, for *Instantaneous*– The first strategy is to ignore the time-window and to execute the same algorithm as if the predictor had given an exact date prediction at time t_0 . The algorithm interrupts the current period (of scheduled length T_R), checkpoints during the interval $[t_0 - C_p, t_0]$, and then returns to regular mode: at time t_0 , it resumes the work needed to complete the interrupted period of the regular mode.

(2) **NOCKPTI**, for *No checkpoint during prediction window*– The second strategy is intended for a short prediction window: instead of ignoring it, we acknowledge it, but make the decision not to checkpoint during it. As in the first strategy, the algorithm interrupts the current period (of scheduled length T_R), and checkpoints during the interval $[t_0 - C_p, t_0]$. But here, we return to regular mode only at time $t_0 + I$, where we resume the work needed to complete the interrupted period of the regular mode. During the whole length of the time-window, we execute work without checkpointing, at the risk of losing work if a fault indeed strikes. But for a small value of I , it may not be worthwhile to checkpoint during the prediction window (if at all possible, since there is no choice if $I < C_p$).

(3) **WITHCKPTI**, for *With checkpoints during prediction window*– The third strategy is intended for a longer prediction window and assumes that $C_p \leq I$: the algorithm interrupts the current period (of scheduled length T_R), and checkpoints during the interval $[t_0 - C_p, t_0]$, but now also decides to take several checkpoints during the prediction window. The period T_P of these checkpoints in proactive mode will presumably be shorter than T_R , to take into account the higher fault probability. In the following, we analytically compute the optimal number of such periods. But we take at least one period here, hence one checkpoint, which implies $C_p \leq I$. We return to regular mode either right after the fault strikes within the time window $[t_0, t_0 + I]$, or at time $t_0 + I$ if no actual fault happens within this window. Then, we resume the work needed to complete the interrupted period of the regular mode. The third strategy is the most complex to describe, and the complete behavior of the corresponding scheduling algorithm is shown in Algorithm 1.

Note that, for all strategies, we insert some additional work for the particular case where there is not enough time to take a checkpoint before entering proactive mode (because a checkpoint for the regular mode is currently on-going).

We account for this work as idle time in the expression of the waste, to ease the analysis. Our expression of the waste is thus an upper bound.

Algorithm 1: WITHCKPTI.

```

1 if fault happens then
2   | After downtime, execute recovery;
3   | Enter regular mode;
4 if in proactive mode for a time greater than or equal to I then
5   | Switch to regular mode
6 if Prediction made with interval  $[t, t + I]$  and prediction taken into
   account then
7   | Let  $t_C$  be the date of the last checkpoint under regular mode to start
   | no later than  $t - C_p$ ;
8   | if  $t_C + C < t - C_p$  then (enough time for an extra checkpoint)
9   |   | Take a checkpoint starting at time  $t - C_p$ 
10  | else (no time for the extra checkpoint)
11  |   | Work in the time interval  $[t_C + C, t]$ 
12  |    $W_{reg} \leftarrow \max(0, t - C_p - (t_C + C))$ ;
13  |   Switch to proactive mode at time  $t$ ;
14 while in regular mode and no predictions are made and no faults happen
   do
15   | Work for a time  $T_R - W_{reg} - C$  and then checkpoint;
16   |  $W_{reg} \leftarrow 0$ ;
17 while in proactive mode and no faults happen do
18   | Work for a time  $T_P - C_p$  and then checkpoint;
```

3.2 Strategy WithCkptI

In this section we evaluate the execution time under heuristic WITHCKPTI. To do so, we partition the whole execution into time intervals defined by the presence or absence of events. An interval starts and ends with either the completion of a checkpoint or of a recovery (after a failure). To ease the analysis, we make a simplifying hypothesis: we assume that *at most* one event, failure or prediction, occurs within any interval of length $T_R + I + C_p$. In particular, this implies that a prediction or an unpredicted fault always take place during the regular mode.

We list below the four types of intervals, and evaluate their respective average length, together with the average work completed during each of them (see Table 1 for a summary):

1. **Two consecutive regular checkpoints with no intermediate events.**
The time elapsed between the completion of the two checkpoints is exactly T_R , and the work done is exactly $T_R - C$.
2. **Unpredicted fault.** Recall that, because of the simplifying hypothesis, the fault happens in regular mode. Because instants where the fault strikes and where the last checkpoint was taken are independent, on average the fault strikes at time $T_R/2$. A downtime of length D and a recovery of length R occur before the interval completes. There is no work done.

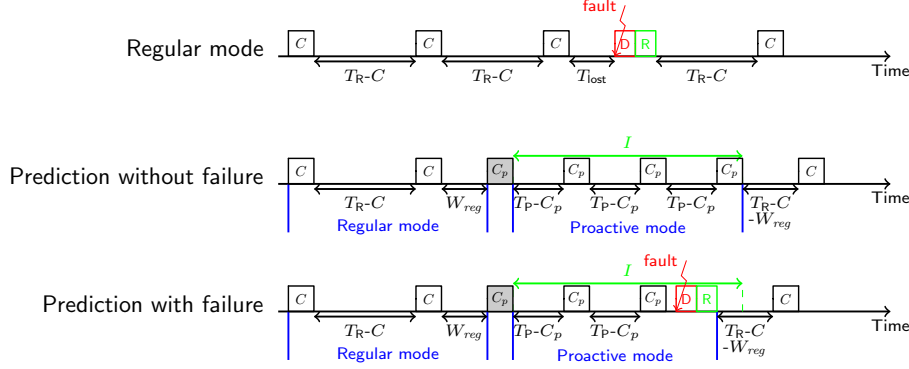


Figure 1: Outline of Algorithm 1 (strategy WITHCKPTI).

3. **False prediction.** Recall that it happens in regular mode. There are two cases:

- (a) **Taken into account.** This happens with probability q . The interval lasts $T_R + C_p + I$, since we take a proactive checkpoint and spend the time I in proactive mode. The work done is $(T_R - C) + (I - \frac{I}{T_P} C_p)$.
- (b) **Not taken into account.** This happens with probability $1 - q$. The interval lasts T_R and the work done is $T_R - C$.

Considering both cases with their probabilities, the average time spent is equal to: $q(T_R + C_p + I) + (1 - q)T_R = T_R + q(C_p + I)$. The average work done is: $q(T_R - C + I - \frac{I}{T_P} C_p) + (1 - q)(T_R - C) = T_R - C + q(I - \frac{I}{T_P} C_p)$.

4. **True prediction.** Recall that it happens in regular mode. There are two cases:

- (a) **Taken into account.** Let $\mathbb{E}_I^{(f)}$ be the average time at which a fault occurs within the prediction window (the time at which the fault strikes is certainly correlated to the starting time of the prediction window; $\mathbb{E}_I^{(f)}$ may not be equal to $I/2$). Up to time $\mathbb{E}_I^{(f)}$, we work and checkpoint in proactive mode, with period T_P . In addition, we take a proactive checkpoint right before the start of the prediction window. Then we spend the time $\mathbb{E}_I^{(f)}$ in proactive mode, and we have a downtime and a recovery. Hence, such an interval lasts $T_R + C_p + \mathbb{E}_I^{(f)} + D + R$ on average. The total work done during the interval is $T_R - C + x(T_P - C_p)$ where x is the expectation of the number of proactive checkpoints successfully taken during the prediction window. Here, $x \approx \frac{\mathbb{E}_I^{(f)}}{T_P} - 1$.
- (b) **Not taken into account.** On average the fault occurs at time $T_R/2$. The time interval has duration $T_R/2 + D + R$, and there is no work done.

Overall the time spent is $q(T_R + C_p + \mathbb{E}_I^{(f)} + D + R) + (1 - q)(T_R/2 + D + R)$,

and the work done is $q(T_R - C + (\frac{\mathbb{E}_I^{(f)}}{T_P} - 1)(T_P - C_p)) + (1 - q)0$.

So far, we have evaluated the length, and the work done, for each of the interval types. We now estimate the expectation of the number of intervals of each type. Consider the intervals defined by an event whose mean time between occurrences is λ . On average, during a time T , there will be T/λ such intervals. Due to

Mode	Number of intervals	Time spent	Work done
(1)	w_1	T_R	$T_R - C$
(2)	$w_2 = \frac{\text{TIME}_{\text{Final}}}{\mu_{NP}}$	$T_R/2 + D + R$	0
(3)	$w_3 = \frac{(1-p)\text{TIME}_{\text{Final}}}{\mu_P}$	$T_R + q(I + C_p)$	$T_R - C + q(I - \frac{I}{T_P}C_p)$
(4)	$w_4 = \frac{p\text{TIME}_{\text{Final}}}{\mu_P}$	$q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q)T_R/2 + D + R$	$q\left(T_R - C + \left(\frac{\mathbb{E}_I^{(f)}}{T_P} - 1\right)(T_P - C_p)\right)$

Table 1: Summary of the different types of interval for WITHCKPTI.

the simplifying hypothesis, intervals of different types never overlap. Table 1 presents the estimation of the number of intervals of each type.

We want to estimate the total execution time. To estimate the time spent within intervals of a given type, we multiply the expectation of the number of intervals of that type by the expectation of the time spent in each of them. Of course, multiplying expectations is correct only if the corresponding random variables are independent. Nevertheless, we hope that this will lead us to a good approximation of the expected execution time. We will assess the quality of the approximation through simulations in Section 4. With our assumptions we have:

$$\begin{aligned} \text{TIME}_{\text{Final}} = w_1 \times T_R + w_2 \left(\frac{T_R}{2} + D + R \right) + w_3 (T_R + q(I + C_p)) \\ + w_4 \left(q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q)\frac{T_R}{2} + D + R \right) \end{aligned} \quad (1)$$

We use the same line of reasoning to compute the overall amount of work done, that must be equal, by definition, to $\text{TIME}_{\text{base}}$, the execution time of the application without any overhead:

$$\begin{aligned} \text{TIME}_{\text{base}} = w_1(T_R - C) + w_2 \times 0 + w_3 \left(T_R - C + q \left(I - \frac{I}{T_P}C_p \right) \right) \\ + w_4 \left(q \left(T_R - C + \left(\frac{\mathbb{E}_I^{(f)}}{T_P} - 1 \right) (T_P - C_p) \right) \right) \end{aligned} \quad (2)$$

This equation gives the value of w_1 as a function of the other parameters. Looking at Equations (1) and (2), and at the values of w_2 , w_3 , and w_4 , we remark that $\text{TIME}_{\text{Final}}$ can be rewritten as a function of q , as follows: $\text{TIME}_{\text{Final}} = \alpha \text{TIME}_{\text{base}} + \beta \text{TIME}_{\text{Final}} + q\gamma \text{TIME}_{\text{Final}}$, that is $\text{TIME}_{\text{Final}} = \frac{\alpha}{1-\beta-q\gamma} \text{TIME}_{\text{base}}$, where neither α , nor β , nor γ depend on q . With a simple differentiation of $\text{TIME}_{\text{Final}}$ with respect to q , we obtain that $\text{TIME}_{\text{Final}}$ is either increasing or decreasing with q , depending on the sign of γ . Consequently, in an optimal solution, either $q = 0$ or $q = 1$. This (somewhat unexpected) conclusion is that the predictor should sometimes be always trusted, and sometimes never, but no in-between value for q will do a better job. Thus we can now focus on the two functions $\text{TIME}_{\text{Final}}$, the one when $q = 0$ ($\text{TIME}_{\text{Final}}^{\{0\}}$), and the one when $q = 1$ ($\text{TIME}_{\text{Final}}^{\{1\}}$).

From Table 1 and Equations (1) and (2), one can easily see that

$$\text{TIME}_{\text{Final}}^{\{0\}} = \frac{T_R}{T_R - C} \text{TIME}_{\text{base}} + \frac{\text{TIME}_{\text{Final}}^{\{0\}}}{\mu} \left(\frac{T_R}{2} + D + R \right), \text{ i.e., that}$$

$$\left(1 - \frac{C}{T_R}\right) \left(1 - \frac{T_R/2 + D + R}{\mu}\right) \text{TIME}_{\text{Final}}^{\{0\}} = \text{TIME}_{\text{base}} \quad (3)$$

This is exactly the equation from [4] in the case of exact-date predictions that are never taken into account (a good sanity check!). When $q = 1$, we have:

$$\begin{aligned} \text{TIME}_{\text{Final}}^{\{1\}} &= \text{TIME}_{\text{base}} \frac{T_R}{T_R - C} \\ &- \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} \frac{T_R}{T_R - C} \left((T_R - C) + (1-p) \left(I - \frac{I}{T_P} C_p \right) + p \left(\frac{\mathbb{E}_I^{(f)}}{T_P} - 1 \right) (T_P - C_p) \right) \\ &+ \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_{NP}} \left(\frac{T_R}{2} + D + R \right) + \frac{(1-p)\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} (T_R + I + C_p) \\ &+ \frac{p\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} \left(T_R + C_p + \mathbb{E}_I^{(f)} + D + R \right) \end{aligned}$$

After a little rewriting we obtain:

$$\begin{aligned} \frac{\text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}^{\{1\}}} &= \frac{r}{p\mu} \left(1 - \frac{C_p}{T_P} \right) \left((1-p)I + p \left(\mathbb{E}_I^{(f)} - T_P \right) \right) \\ &+ \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1-r)p \frac{T_R}{2} + r \left((1-p)I + p\mathbb{E}_I^{(f)} \right) \right) \right) \end{aligned}$$

Finally, the waste is equal by definition to $\frac{\text{TIME}_{\text{Final}} - \text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}}$. Therefore, we have:

$$\begin{aligned} \text{WASTE} &= 1 - \frac{r}{p\mu} \left(1 - \frac{C_p}{T_P} \right) \left((1-p)I + p \left(\mathbb{E}_I^{(f)} - T_P \right) \right) \\ &- \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1-r)p \frac{T_R}{2} + r \left((1-p)I + p\mathbb{E}_I^{(f)} \right) \right) \right) \quad (4) \end{aligned}$$

Waste minimization

When $q = 0$, the optimal period can readily be computed from Equation (3) and we derive that the optimal period is $\sqrt{2(\mu - (D + R))C}$. This defines a periodic policy we call RFO, for Refined First-Order approximation. We now minimize the waste of the strategy where $q = 1$. In order to compute the optimal value for T_P , we identify the fraction of the waste in Equation (4) that depends on T_P . We can rewrite Equation (4) as:

$$\text{WASTE}^{\{1\}} = \alpha + \frac{r}{p\mu} \left(\left((1-p)I + p\mathbb{E}_I^{(f)} \right) \frac{C_p}{T_P} + pT_P \right)$$

where α does not depend on T_P . The waste is thus minimized when T_P is equal to $T_P^{\text{extr}} = \sqrt{\frac{((1-p)I + p\mathbb{E}_I^{(f)})C_p}{p}}$. Note that we always have to enforce that T_P^{extr} is larger than C_p and does not exceed I , and we may have to round its values accordingly in some extreme cases.

In order to compute the optimal value for T_R , we identify the fraction of the waste in Equation (4) that depends on T_R . We can rewrite Equation (4) as:

$$\text{WASTE}^{\{1\}} = \beta + \frac{C}{T_R} \left(1 - \frac{1}{p\mu} \left(p(D+R) + r \left(C_p + (1-p)I + p\mathbb{E}_I^{(f)} \right) \right) \right) + \frac{1-r}{\mu} \frac{T_R}{2} \quad (5)$$

where β does not depend on T_R because T_P^{opt} does not depend on T_R . Therefore, $\text{WASTE}^{\{1\}}$ is minimized when T_R is equal to

$$T_R^{\text{extr}} = \sqrt{\frac{2C \left(p\mu - \left(p(D+R) + r \left(C_p + \left((1-p)I + p\mathbb{E}_I^{(f)} \right) \right) \right) \right)}{p(1-r)}} \quad (6)$$

Recall that we must always enforce that T_R^{extr} is always greater than C .

One can note that when $r = 0$, this means that none of the prediction predicts an actual fault, and we obtain the same period than without a predictor. Finally, if we assume that, on average, fault strikes at the middle of the prediction window, i.e., $\mathbb{E}_I^{(f)} = \frac{I}{2}$, we obtain simplified values:

$$T_P^{\text{extr}} = \sqrt{\frac{(2-p)IC_p}{p}} \text{ and } T_R^{\text{extr}} = \sqrt{\frac{2C \left(p\mu - \left(p(D+R) + r \left(C_p + \left(1 - \frac{p}{2} \right) I \right) \right) \right)}{p(1-r)}}$$

3.3 Strategy NoCkptI

In this section we evaluate the execution time under heuristic NOCKPTI. The analysis is rather similar to that of WITHCKPTI, the only differences being, obviously, in the presence of true and false predictions:

3. **False prediction.** There are two cases:

- (a) **Taken into account.** This happens with probability q . The interval lasts $T_R + C_p + I$, since we take a proactive checkpoint and spend the time I in proactive mode (here, working without checkpointing). The work done is $(T_R - C) + I$.
- (b) **Not taken into account.** This happens with probability $1 - q$. The interval lasts T_R and the work done is $T_R - C$.

Considering both cases with their probabilities, the average time spent is equal to: $q(T_R + C_p + I) + (1 - q)T_R = T_R + q(C_p + I)$. The average work done is: $q(T_R - C + I) + (1 - q)(T_R - C) = T_R - C + qI$.

4. **True prediction.** There are two cases:

- (a) **Taken into account.** Let $\mathbb{E}_I^{(f)}$ be the average time at which a fault occurs within the prediction window. We take a proactive checkpoint right before the start of the prediction window. Then we spend the time $\mathbb{E}_I^{(f)}$ in proactive mode working without checkpointing, and we have a downtime and a recovery. Hence, such an interval lasts $T_R + C_p + \mathbb{E}_I^{(f)} + D + R$ on average. The total work done during the interval is $T_R - C$.
- (b) **Not taken into account.** On average the fault occurs at time $T_R/2$. The time interval has duration $T_R/2 + D + R$, and there is no work done.

Mode	Number of intervals	Time spent	Work done
(1)	w_1	T_R	$T_R - C$
(2)	$w_2 = \frac{\text{TIME}_{\text{Final}}}{\mu_{NP}}$	$T_R/2 + D + R$	0
(3)	$w_3 = \frac{(1-p)\text{TIME}_{\text{Final}}}{\mu_P}$	$T_R + q(I + C_p)$	$T_R - C + qI$
(4)	$w_4 = \frac{p\text{TIME}_{\text{Final}}}{\mu_P}$	$q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q)T_R/2 + D + R$	$q(T_R - C)$

Table 2: Summary of the different types of interval for NoCKPTI.

Overall the time spent is $q(T_R + C_p + \mathbb{E}_I^{(f)} + D + R) + (1-q)(T_R/2 + D + R)$, and the work done is $q(T_R - C) + (1-q)0$.

So far, we have evaluated the length, and the work done, for each of the interval types. We now estimate the expectation of the number of intervals of each type as we did for WITHCKPTI. Table 2 presents the estimation of the number of intervals of each type.

We estimate the total execution time as for WITHCKPTI. The formula is the exact same function of w_1 , w_2 , w_3 , and w_4 (but the values of there four parameters will change as the average work done during some of the types of intervals changes):

$$\begin{aligned} \text{TIME}_{\text{Final}} = w_1 \times T_R + w_2 \left(\frac{T_R}{2} + D + R \right) + w_3 (T_R + q(I + C_p)) \\ + w_4 \left(q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q)\frac{T_R}{2} + D + R \right) \end{aligned} \quad (7)$$

We use the same line of reasoning as previously to compute the overall amount of work done:

$$\text{TIME}_{\text{base}} = w_1(T_R - C) + w_2 \times 0 + w_3 (T_R - C + qI) + w_4 (q(T_R - C)) \quad (8)$$

This equation gives the value of w_1 as a function of the other parameters. As for WITHCKPTI, one can easily show that in an optimal solution, either $q = 0$ or $q = 1$. Thus we can now focus on the two functions $\text{TIME}_{\text{Final}}$, the one when $q = 0$ ($\text{TIME}_{\text{Final}}^{\{0\}}$), and the one when $q = 1$ ($\text{TIME}_{\text{Final}}^{\{1\}}$).

From Table 2 and Equations (7) and (8), one can easily see that

$$\begin{aligned} \text{TIME}_{\text{Final}}^{\{0\}} = \frac{T_R}{T_R - C} \text{TIME}_{\text{base}} + \frac{\text{TIME}_{\text{Final}}^{\{0\}}}{\mu} \left(\frac{T_R}{2} + D + R \right), \text{ i.e., that} \\ \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{T_R/2 + D + R}{\mu} \right) \text{TIME}_{\text{Final}}^{\{0\}} = \text{TIME}_{\text{base}} \end{aligned} \quad (9)$$

This is exactly the equation from [4] in the case of exact-date predictions that are never taken into account, what we had already retrieved with WITHCKPTI

(same sanity check!). When $q = 1$, we have:

$$\begin{aligned} \text{TIME}_{\text{Final}}^{\{1\}} &= \text{TIME}_{\text{base}} \frac{T_R}{T_R - C} - \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} \frac{T_R}{T_R - C} ((T_R - C) + (1 - p)I) \\ &+ \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_{NP}} \left(\frac{T_R}{2} + D + R \right) + \frac{(1 - p)\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} (T_R + I + C_p) \\ &+ \frac{p\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} (T_R + C_p + \mathbb{E}_I^{(f)} + D + R) \end{aligned}$$

After a little rewriting we obtain:

$$\begin{aligned} \frac{\text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}^{\{1\}}} &= \frac{r}{p\mu} (1 - p)I \\ &+ \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1 - r)p\frac{T_R}{2} + r((1 - p)I + p\mathbb{E}_I^{(f)}) \right) \right) \end{aligned}$$

Finally, the waste is equal by definition to $\frac{\text{TIME}_{\text{Final}} - \text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}}$. Therefore, we have:

$$\begin{aligned} \text{WASTE} &= 1 - \frac{r}{p\mu} (1 - p)I \\ &- \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1 - r)p\frac{T_R}{2} + r((1 - p)I + p\mathbb{E}_I^{(f)}) \right) \right) \end{aligned} \quad (10)$$

Waste minimization

When $q = 0$, the optimal value for T_R is obviously the same than the one we computed for WITHCKPTI in the case $q = 0$. We now minimize the waste of the strategy where $q = 1$. In order to compute the optimal value for T_R , we identify the fraction of the waste in Equation (10) that depends on T_R . We can rewrite Equation (10) as:

$$\text{WASTE}^{\{1\}} = \beta + \frac{C}{T_R} \left(1 - \frac{1}{p\mu} \left(p(D + R) + r(C_p + (1 - p)I + p\mathbb{E}_I^{(f)}) \right) \right) + \frac{1 - r}{\mu} \frac{T_R}{2}$$

where β does not depend on T_R . This equation is identical to Equation (5) and therefore the value of T_R that minimizes the waste is T_R^{extr} , the value given by Equation (6).

3.4 Strategy Instant

In this section we evaluate the execution time under heuristic INSTANT. The analysis is very similar to that of NOCKPTI. Indeed, we only focus to the differences between the performance of INSTANT and WITHCKPTI. The differences happening, obviously, only in the presence of true and false predictions:

3. False prediction. There are two cases:

Mode	Number of intervals	Time spent	Work done
(1)	w_1	T_R	$T_R - C$
(2)	$w_2 = \frac{\text{TIME}_{\text{Final}}}{\mu_{NP}}$	$T_R/2 + D + R$	0
(3)	$w_3 = \frac{(1-p)\text{TIME}_{\text{Final}}}{\mu_P}$	$T_R + qC_p$	$T_R - C$
(4)	$w_4 = \frac{p\text{TIME}_{\text{Final}}}{\mu_P}$	$q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q)T_R/2 + D + R$	$q(T_R - C)$

Table 3: Summary of the different types of interval for INSTANT.

- (a) **Taken into account.** This happens with probability q . The interval lasts $T_R + C_p$, since we fallback to regular mode as soon as the proactive checkpoints completes. The work done is $T_R - C$.
- (b) **Not taken into account.** This happens with probability $1 - q$. The interval lasts T_R and the work done is $T_R - C$.

Considering both cases with their probabilities, the average time spent is equal to: $q(T_R + C_p) + (1 - q)T_R = T_R + qC_p$. The average work done is: $q(T_R - C) + (1 - q)(T_R - C) = T_R - C$.

4. **True prediction.** There are two cases:

- (a) **Taken into account.** Let $\mathbb{E}_I^{(f)}$ be the average time at which a fault occurs within the prediction window. We take a proactive checkpoint right before the start of the prediction window. Then we fallback to the regular mode. After a time $\mathbb{E}_I^{(f)}$ the fault strikes. Depending on the size I of the prediction window, and of when the prediction started after the completion of the last regular checkpoint three scenarios can happen. Either the fault strikes while the heuristic is still trying to complete the work of size $T_R - C$, or it strikes while the heuristic is trying to take the regular checkpoint after that work, or it strikes after that regular checkpoint was completed. We overestimate the time lost by assuming that we are in one of the two former cases, because these are the cases that maximizes the amount of work destroyed by a strike. (In some way, this is equivalent to assuming that I is very small with respect to T_R .) The predicted fault and the completion time of the last regular checkpoint are independent events. Therefore, on average the fault strikes at time $T_R/2$. After the fault strikes, the downtime and the recovery we complete the period struck by the fault. Then, the interval lasts $T_R + C_p + \mathbb{E}_I^{(f)} + D + R$ on average. The total work done during the interval is $T_R - C$.
- (b) **Not taken into account.** On average the fault occurs at time $T_R/2$. The time interval has duration $T_R/2 + D + R$, and there is no work done.

Overall the time spent is $q(T_R + C_p + \mathbb{E}_I^{(f)} + D + R) + (1 - q)(T_R/2 + D + R)$, and the work done is $q(T_R - C)$.

So far, we have evaluated the length, and the work done, for each of the interval types. We estimate the expectation of the number of intervals of each type as we did for WITHCKPTI and for NOCKPTI. Table 3 presents the estimation of the number of intervals of each type.

We estimate the total execution time as for WITHCKPTI and NOCKPTI:

$$\begin{aligned} \text{TIME}_{\text{Final}} = w_1 \times T_R + w_2 \left(\frac{T_R}{2} + D + R \right) + w_3 (T_R + qC_p) \\ + w_4 \left(q(T_R + \mathbb{E}_I^{(f)} + C_p) + (1-q) \frac{T_R}{2} + D + R \right) \end{aligned} \quad (11)$$

We use the same line of reasoning as previously to compute the overall amount of work done:

$$\text{TIME}_{\text{base}} = w_1(T_R - C) + w_2 \times 0 + w_3(T_R - C) + w_4(q(T_R - C)) \quad (12)$$

This equation gives the value of w_1 as a function of the other parameters. As with WITHCKPTI and NOCKPTI, one can easily show that in an optimal solution, either $q = 0$ or $q = 1$. Thus we can now focus on the two functions $\text{TIME}_{\text{Final}}$, the one when $q = 0$ ($\text{TIME}_{\text{Final}}^{\{0\}}$), and the one when $q = 1$ ($\text{TIME}_{\text{Final}}^{\{1\}}$).

From Table 3 and Equations (11) and (12), one can easily see that

$$\begin{aligned} \text{TIME}_{\text{Final}}^{\{0\}} = \frac{T_R}{T_R - C} \text{TIME}_{\text{base}} + \frac{\text{TIME}_{\text{Final}}^{\{0\}}}{\mu} \left(\frac{T_R}{2} + D + R \right), \text{ i.e., that} \\ \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{T_R/2 + D + R}{\mu} \right) \text{TIME}_{\text{Final}}^{\{0\}} = \text{TIME}_{\text{base}} \end{aligned} \quad (13)$$

This is exactly the equation from [4] in the case of exact-date predictions that are never taken into account, what we had already remarked with WITHCKPTI and NOCKPTI (yet another good sanity check!). When $q = 1$, we have:

$$\begin{aligned} \text{TIME}_{\text{Final}}^{\{1\}} = \text{TIME}_{\text{base}} \frac{T_R}{T_R - C} - \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} T_R + \frac{\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_{NP}} \left(\frac{T_R}{2} + D + R \right) \\ + \frac{(1-p)\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} (T_R + C_p) + \frac{p\text{TIME}_{\text{Final}}^{\{1\}}}{\mu_P} (T_R + C_p + \mathbb{E}_I^{(f)} + D + R) \end{aligned}$$

After a little rewriting we obtain:

$$\frac{\text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}^{\{1\}}} = \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1-r)p \frac{T_R}{2} + pr\mathbb{E}_I^{(f)} \right) \right)$$

Finally, the waste is equal by definition to $\frac{\text{TIME}_{\text{Final}} - \text{TIME}_{\text{base}}}{\text{TIME}_{\text{Final}}}$. Therefore, we have:

$$\text{WASTE} = 1 - \left(1 - \frac{C}{T_R} \right) \left(1 - \frac{1}{p\mu} \left(p(D + R) + rC_p + (1-r)p \frac{T_R}{2} + pr\mathbb{E}_I^{(f)} \right) \right) \quad (14)$$

Waste minimization

When $q = 0$, the optimal value for T_R is obviously the same than the one we computed for WITHCKPTI and for NOCKPTI in the case $q = 0$. We now

minimize the waste of the strategy where $q = 1$. In order to compute the optimal value for T_R , we identify the fraction of the waste in Equation (14) that depends on T_R . We can rewrite Equation (14) as:

$$\text{WASTE}^{\{1\}} = \beta + \frac{C}{T_R} \left(1 - \frac{1}{p\mu} \left(p(D+R) + rC_p + pr\mathbb{E}_I^{(f)} \right) \right) + \frac{1-r}{\mu} \frac{T_R}{2}$$

where β does not depend on T_R . Therefore, the value of T_R that minimizes the waste is T_R^{extr} , where

$$T_R^{\text{extr}} = \sqrt{\frac{2C \left(p\mu - \left(p(D+R) + rC_p + pr\mathbb{E}_I^{(f)} \right) \right)}{p(1-r)}}$$

Again, recall that we must always enforce that T_R^{extr} is always greater than C . Finally, if we assume that, on average, fault strikes at the middle of the prediction window, i.e., $\mathbb{E}_I^{(f)} = \frac{I}{2}$, we have:

$$T_R^{\text{extr}} = \sqrt{\frac{2C \left(p\mu - \left(p(D+R) + rC_p + pr\frac{I}{2} \right) \right)}{p(1-r)}}$$

4 Simulation results

We start by presenting the simulation framework (Section 4.1). Then we report results using the characteristics of two fault predictors from the literature (Section 4.2).

4.1 Simulation framework

In order to validate the model, we have instantiated it with several scenarios. The experiments use parameters that are representative of current and forthcoming large-scale platforms [11, 12]. We take $C = R = 600$ seconds, and $D = 60$ seconds. We consider three scenarios where proactive checkpoints are (i) exactly as expensive as periodic checkpoints ($C_p = C$); (ii) ten times cheaper ($C_p = 0.1C$); and (iii) two times more expensive ($C_p = 2C$). The individual (processor) MTBF is $\mu_{\text{ind}} = 125$ years, and the total number of processors N varies from $N = 2^{16} = 16,384$ to $N = 2^{19} = 524,288$, so that the platform MTBF μ varies from $\mu = 4,010$ min (about 2.8 days) down to $\mu = 125$ min (about 2 hours). For instance the Jaguar platform, with $N = 45,208$ processors, is reported to have experienced about one fault per day [13], which leads to $\mu_{\text{ind}} = \frac{45,208}{365} \approx 125$ years. The application size is set to $\text{TIME}_{\text{base}} = 10,000$ years/ N .

We use Maple to analytically compute and plot the optimal value of the waste for the three prediction-aware policies, `INSTANT`, `NOCKPTI`, and `WITHCKPTI`, for the prediction-ignoring policy `RFO` (corresponding to the case $q = 0$), and for the reference heuristic `DALY` (Daly's [2] periodic policy). In order to check the accuracy of our model, we have compared the analytical results with results obtained with a discrete-event simulator. The simulation engine generates a random trace of faults, parameterized either by an Exponential fault distribution or by Weibull distribution laws with shape parameter 0.5 or 0.7. Note

that Exponential faults are widely used for theoretical studies, while Weibull faults are representative of the behavior of real-world platforms [14, 15, 16]. In both cases, the distribution is scaled so that its expectation corresponds to the platform MTBF μ . With probability r , we decide if a fault is predicted or not. The simulation engine also generates a random trace of false predictions, whose distribution is identical to that of the first trace (in Figures 8 through 13, we also consider the case where false predictions are generated according to a uniform distribution; results are quite similar). This second distribution is scaled so that its expectation is equal to $\frac{\mu_P}{1-p} = \frac{p\mu}{r(1-p)}$, the inter-arrival time of false predictions. Finally, both traces are merged to produce the final trace including all events (true predictions, false predictions, and non predicted faults). Each reported value is the average over 100 randomly generated instances.

In the simulations, we compare the five checkpointing strategies listed above. To assess the quality of each strategy, we compare it with its BESTPERIOD counterpart, defined as the same strategy but using the best possible period T_R . This latter period is computed via a brute-force numerical search for the optimal period. Altogether, there are four BESTPERIOD heuristics, one for each of the three variants with prediction, and one for the case where we ignore predictions, which corresponds to both DALY and RFO. Altogether we have a rich set of nine heuristics, which enables us to comprehensively assess the actual quality of the proposed strategies. Note that for computer algebra plots, obviously we do not need BESTPERIOD heuristics, since each period is already chosen optimally from the equations.

We experiment with two predictors from the literature: one accurate predictor with high recall and precision [5], namely with $p = 0.82$ and $r = 0.85$, and another predictor with more limited recall and precision [10], namely with $p = 0.4$ and $r = 0.7$. In both cases, we use five different prediction windows, of size $I = 300, 600, 900, 1200$, and 3000 seconds. Figures 2 through 7 show the average waste degradation of the nine heuristics for both predictors, as a function of the number of processors N . We draw the plots as a function of the number of processors N rather than of the platform MTBF $\mu = \mu_{ind}/N$, because it is more natural to see the waste increase with larger platforms; however, this work is agnostic of the granularity of the processors and intrinsically focuses on the impact of the MTBF on the waste.

4.2 Analysis of the results

We start with a preliminary remark: when the graphs for INSTANT and WITH-CKPTI cannot be seen in the figures, this is because their performance is identical to that of NOCKPTI, and their respective graphs are superposed.

We first compare the analytical results, plotted by the Maple curves, to the simulations results. There is a good correspondence between the analytical curves and the simulations, especially those using an Exponential distribution of failures. However, the larger the platform (or the smaller the MTBF), the less realistic our assumption that no two events happen during an interval of length $T_R + I + C_p$, and the analytical models become less accurate for prediction-aware heuristics. Therefore, the analytical results are overly pessimistic in the most failure-prone platforms. Also, recall that an exponential law is a Weibull law of shape parameter 1. Therefore, the further the distribution of failures is from an exponential law, the larger the difference between analytical results and

simulated ones. However, in all cases, the analytical results are able to predict the general *trends*.

A second assessment of the quality of our analysis comes from the BESTPERIOD variants of our heuristics. When predictions are not taken into account, DALY, and to a lesser extent RFO, are not close to the optimal period given by BESTPERIOD (a similar observation was made in [17]). This gap increases when the distribution is further apart from an Exponential distribution. However, prediction-aware heuristics are very close to BESTPERIOD in almost all configurations. The only exception is with heuristics INSTANT when $C_p = 2C$, the total number of processors N is equal to either 2^{18} and 2^{19} , and I is large. However, when $I = 3000$ and $N = 2^{19}$, the platform MTBF is approximately equal to $6C_p$ which renders our hypothesis and analysis invalid. The difference in this case between INSTANT and its BESTPERIOD should therefore not come as a surprise.

To better understand why close-to-optimal periods are obtained by prediction-aware heuristics (while this is not the case without predictions), we plot the waste as a function of the period T_R for RFO and the prediction-aware heuristics (Figures 14 through 17). On these figures one can see that, whatever the configuration, periodic checkpointing policies (ignoring predictions) have well-defined global optimum. (One should nevertheless remark that the performance is almost constant in the neighborhood of the optimal period which explains why policies using different periods can obtain in practice similar performance, as in [18].) For prediction-aware heuristics, however, the behavior is quite different and two scenarios are possible. In the first one, once the optimum is reached, the waste very slowly increases to reach an asymptotic value which is close to the optimum waste (e.g., when the platform MTBF is large and failures follow an exponential distribution). Therefore, any period chosen close to the optimal one, or greater than it, will deliver good quality performance. In the second scenario, the waste decreases until the period becomes larger than the application size, and the waste stays constant. In other words, in these configurations, periodic checkpointing is unnecessary, only proactive actions matter! This striking result can be explained as follows: a significant fraction of the failures are predicted, and thus taken care of, by proactive checkpoints. The impact of unpredicted failures is mitigated by the proactive measures taken for false predictions. To further mitigate the impact of unpredicted faults, the period T_R should be significantly shorter than the mean-time between proactive checkpoints, which would induce a lot of waste due to unnecessary checkpoints if the mean-time between unpredicted faults is large with respect to the mean-time between predictions. This greatly restrict the scenarios for which the periodic checkpointing can lead to a significant decrease of the waste.

When the prediction window I is shorter than the duration C_p of a proactive checkpoint, there is no difference between NOCKPTI and WITHCKPTI. When I is small but greater than C_p (say, when I is around $2C_p$), WITHCKPTI spends most of the prediction window taking a proactive checkpoint and NOCKPTI is more efficient. When I becomes “large” with respect to C_p , WITHCKPTI can become more efficient than NOCKPTI, but becomes significantly more efficient only if the proactive checkpoints are significantly shorter than regular ones. INSTANT can hardly be seen in the graphs as its performance is most of the time equivalent to that of NOCKPTI.

Figures 18 through 21 show the influence of the size of the prediction window

I on the performance of the heuristics. As expected, the smaller the prediction window, the more efficient the prediction-aware heuristics. Also, the smaller the number of processors (or the larger the platform MTBF), the larger the impact of the size of the prediction window. A surprising result is that taking prediction into account is not always beneficial! The analytical results predict that prediction-aware heuristics would achieve worse performance than periodic policies in our settings, as soon as the platform includes 2^{18} processors. In simulations, results are not so extreme. For the largest platforms considered, using predictions has almost no impact on performance. But when the prediction window is very large, taking predictions into account can indeed be detrimental. These observations can be explained as follows. When the platform includes 2^{19} processors, the platform MTBF is equal to 7500 s. Therefore, any interval of duration 3000 has a 40% chance to include a failure: a prediction window of 3000 is not very informative, unless the precision and recall of the predictor are almost equal to 1 (which is never the case in practice). Since the predictor brings almost no knowledge, trusting it may be detrimental. When comparing the performance of, say, NOCKPTI for the two predictors, one can see that when failures follow a Weibull distribution with shape parameter $k = 0.7$, $I = 600$, and $N = 2^{18}$, NOCKPTI achieves better performance than RFO when $r = 0.85$ and $p = 0.82$, but worse when $p = 0.4$ and $r = 0.7$. The latter predictor generates more false predictions—each one inducing an unnecessary proactive checkpoint—and misses more actual failures—each one destroying some work. The drawbacks of trusting the predictor outweigh the advantages. If failures are few and apart, almost any predictor will be beneficial. When the platform MTBF is small with respect to the cost of proactive checkpoints, only almost perfect predictors will be worth using. For each set of predictor characteristics, there is a threshold for the platform MTBF under which predictions will be useless or detrimental, but above which predictions will be beneficial.

In order to compare the impact of the heuristics ignoring predictions to those using them, we report job execution times in Table 4. For the strategies with prediction, we compute the gain (expressed in percentage) over DALY, the reference strategy without prediction. We first remark that RFO achieves lower makespans than DALY with gains ranging from 1% with 2^{16} processors to 18% with 2^{19} processors. Overall, the gain due to the predictions decreases when the size of the prediction window increases, and increases with the platform size. This gain is obviously closely related to the characteristics of the predictor.

When $I = 300$, the three strategies are identical. When I increases, NOCKPTI achieves slightly better results than INSTANT. For low values of I , WITHCKPTI is the worst prediction-aware heuristics. But when I becomes large and if the predictor is efficient, then WITHCKPTI becomes the heuristics of choice ($I = 3000$, $p = 0.82$, and $r = 0.85$).

The reductions in the application executions times due to the predictor can be very significant. With $p = 0.85$ and $r = 0.82$ and $I = 3000$, we save 25% of the total time with $N = 2^{19}$, and 13% with $N = 2^{16}$ using strategy WITHCKPTI. With $I = 300$, we save up to 45% with $N = 2^{19}$, and 18% with $N = 2^{16}$ using any strategy (though NOCKPTI is slightly better than INSTANT). Then, with $p = 0.4$ and $r = 0.7$, we still save 33% of the execution time when $I = 300$ and $N = 2^{19}$, and 14% with $N = 2^{16}$. The gain gets smaller with $I = 3000$ and $N = 2^{16}$ but remains non negligible since we can save 8%. When $I = 3000$ and $N = 2^{19}$, however, the best solution is to ignore predictions and simply

	$I = 300$ s		$I = 1200$ s		$I = 3000$ s	
	2^{16} procs	2^{19} procs	2^{16} procs	2^{19} procs	2^{16} procs	2^{19} procs
DALY	81.3	31.0	81.3	31.0	81.3	31.0
RFO	80.2 (1%)	25.5 (18%)	80.2 (1%)	25.5 (18%)	80.2 (1%)	25.5 (18%)
$p = 0.82, r = 0.85$						
NoCKPTI	66.4 (18%)	17.0 (45%)	67.9 (16%)	20.2 (35%)	71.0 (13%)	24.7 (20%)
WithCKPTI	66.4 (18%)	17.0 (45%)	68.3 (16%)	20.6 (33%)	70.6 (13%)	23.1 (25%)
INSTANT	66.5 (18%)	17.0 (45%)	68.0 (16%)	20.3 (34%)	70.9 (13%)	24.1 (22%)
$p = 0.4, r = 0.7$						
NoCKPTI	70.2 (14%)	20.6 (33%)	71.8 (12%)	24.2 (22%)	75.0 (8%)	28.7 (7%)
WithCKPTI	70.2 (14%)	20.6 (33%)	73.6 (9%)	25.5 (18%)	75.1 (8%)	26.6 (14%)
INSTANT	70.3 (13%)	20.9 (33%)	72.0 (11%)	24.6 (21%)	75.0 (8%)	27.7 (11%)

Table 4: Job execution times (in days) under the different checkpointing policies, when failures follow a Weibull distribution of shape parameter 0.7. Gains are reported with respect to DALY.

	$I = 300$ s		$I = 1200$ s		$I = 3000$ s	
	2^{16} procs	2^{19} procs	2^{16} procs	2^{19} procs	2^{16} procs	2^{19} procs
DALY	125.7	185.0	125.7	185.0	125.7	185.0
RFO	120.1 (4%)	114.8 (38%)	120.1 (4%)	114.8 (38%)	120.1 (4%)	114.8 (38%)
$p = 0.82, r = 0.85$						
NoCKPTI	77.4 (38%)	44.9 (76%)	81.8 (35%)	60.7 (67%)	90.0 (28%)	71.5 (61%)
WithCKPTI	77.4 (38%)	44.9 (76%)	83.6 (33%)	64.4 (65%)	89.8 (29%)	66.2 (64%)
INSTANT	77.4 (38%)	45.2 (76%)	82.0 (35%)	60.8 (67%)	89.7 (29%)	70.6 (62%)
$p = 0.4, r = 0.7$						
NoCKPTI	84.4 (33%)	58.3 (68%)	89.1 (29%)	76.8 (58%)	97.9 (22%)	83.7 (55%)
WithCKPTI	84.4 (33%)	58.3 (68%)	93.8 (25%)	75.4 (59%)	97.8 (22%)	77.7 (58%)
INSTANT	84.5 (33%)	59.6 (68%)	89.4 (29%)	76.64 (58%)	97.7 (22%)	81.9 (56%)

Table 5: Job execution times (in days) under the different checkpointing policies, when failures follow a Weibull distribution of shape parameter 0.5. Gains are reported with respect to DALY.

use RFO (we fall-back to the case $q = 0$). If we now consider a Weibull law with shape parameter 0.5 instead of 0.7, keeping all other parameters identical ($I = 3000$, $N = 2^{19}$, $p = 0.4$ and $r = 0.7$), then the heuristics of choice is WithCKPTI and the gain with respect to DALY is 57.9%.

5 Related work

Considerable research has been conducted on fault prediction using different models (system log analysis [5], event-driven approach [19, 5, 10], support vector machines [6, 20]), nearest neighbors [6], ...). In this section we give a brief overview of the results obtained by predictors. We focus on their results rather than on their methods of prediction.

The authors of [10] introduce the *lead time*, that is the time between the prediction and the actual fault. This time should be sufficient to take proactive actions. They are also able to give the location of the fault. While this has a negative impact on the precision (see the low value of p in Table 6), they state that it has a positive impact on the checkpointing time (from 1500 seconds to 120 seconds). The authors of [5] also consider a lead time, and introduce a *prediction window* when the predicted fault should happen. The authors of [6] study the impact of different prediction techniques with different prediction

Paper	Lead Time	Precision	Recall	Prediction Window
[10]	300 s	40 %	70%	-
[10]	600 s	35 %	60%	-
[5]	2h	64.8 %	65.2%	yes (size unknown)
[5]	0 min	82.3 %	85.4 %	yes (size unknown)
[19]	32 s	93 %	43 %	-
[20]	NA	70 %	75 %	-
[6]	NA	20 %	30 %	1h
[6]	NA	30 %	75 %	4h
[6]	NA	40 %	90 %	6h
[6]	NA	50 %	30 %	6h
[6]	NA	60 %	85%	12h

Table 6: Comparative study of different parameters returned by some predictors.

window sizes. They also consider a lead time, but do not state its value. These two latter studies motivate this work, even though [5] does not provide the size of their prediction window.

Unfortunately, much of the work done on prediction does not provide information that could be really useful for the design of efficient algorithms. These informations are those stated above, namely the lead time and the size of the prediction window, but other information that could be useful would be: (i) the distribution of the faults in the prediction window; (ii) the precision as a function of the recall (see our analysis); and (iii) the precision and recall as functions of the prediction window (what happens with a larger prediction window).

While many studies on fault prediction focus on the conception of the predictor, most of them consider that the proactive action should simply be a checkpoint or a migration right in time before the fault. However, in their paper [21], Li et al. consider the mathematical problem to determine when and how to migrate. In order to be able to use migration, they stated that at every time, 2% of the resources are available. This allowed them to conceive a Knapsack-based heuristic. Thanks to their algorithm, they were able to save 30% of the execution time compared to an heuristic that does not take the reliability into account, with a precision and recall of 70%, and with a maximum load of 0.7.

In the simpler case where predictions are exact-date predictions, Gainaru et al [3] have shown that the optimal checkpointing period becomes $T_{\text{opt}} = \sqrt{\frac{2\mu C}{1-r}}$, but their analysis is valid only if μ is very large in front of the other parameters. Our previous work [4] has refined the results of [3], focusing on a more accurate analysis of fault prediction with exact dates, and providing a detailed study on the impact of recall and precision on the waste. As shown in Section 3, the analysis of the waste is dramatically more complicated when using prediction windows than when using exact-date predictions. To the best of our knowledge, this work is the first to focus on the mathematical aspect of fault prediction with prediction windows, and to provide a model and a detailed analysis of the waste due to all three types of events (true and false predictions and unpredicted failures).

6 Conclusion

In this work, we have studied the impact of prediction windows on checkpointing strategies. We have designed several heuristics that decide whether to trust these predictions, and when it is worth taking preventive checkpoints. We have been able to derive a comprehensive set of results and conclusions:

- We have introduced an analytical model to capture the waste incurred by each strategy, and provided for each optimization problem a closed-form formula giving its optimal solution. Contrarily to the cases without prediction, or with exact-date predictions, the computation of the waste requires a sophisticated analysis of the various events, including the time spent irregular or proactive modes.
- The simulations fully validate the model, and the brute-force computation of the optimal period guarantees that our prediction-aware strategies are always very close to the optimal. This holds true both for Exponential and Weibull failure distributions.
- The model is quite accurate and its validity goes beyond the conservative assumption that requires a single event per time interval; even more surprising, the accuracy of the model for prediction-aware strategies is much better than for the case without predictions, where DALY can be far from the optimal period in the case of Weibull failure distributions.
- Both the analytical computations and the simulations enable to characterize when prediction is useful, and which strategy performs better, given the key parameters of the system: recall r , precision p , size of the prediction window I , size of proactive checkpoints C_p versus regular checkpoints C , and platform MTBF μ .

Altogether, the analytical model and the comprehensive results provided in this work enable to fully assess the impact of fault prediction with time-windows on optimal checkpointing strategies. Future work will be devoted to refine the assessment of the usefulness of prediction with trace-based failure and prediction logs from current large-scale supercomputers.

Acknowledgments. The authors are with Université de Lyon, France. Y. Robert is with the Institut Universitaire de France. This work was supported in part by the ANR *RESCUE* project.

References

- [1] Young, J.W.: A first order approximation to the optimum checkpoint interval. *Comm. of the ACM* **17**(9) (1974) 530–531
- [2] Daly, J.T.: A higher order estimate of the optimum checkpoint interval for restart dumps. *FGCS* **22**(3) (2004) 303–312
- [3] Gainaru, A., Cappello, F., Kramer, W., Snir, M.: Fault prediction under the microscope - a closer look into hpc systems. In: *SC’12*. (2012)
- [4] Aupy, G., Robert, Y., Vivien, F., Zaidouni, D.: Checkpointing algorithms and fault prediction. Research report RR-8237, INRIA (February 2013)

- [5] Yu, L., Zheng, Z., Lan, Z., Coghlan, S.: Practical online failure prediction for Blue Gene/P: Period-based vs event-driven. In: DSN-Workshops. (2011)
- [6] Liang, Y., Zhang, Y., Xiong, H., Sahoo, R.K.: Failure prediction in IBM BlueGene/L event logs. In: ICDM. (2007) 583–588
- [7] Kolettis, N., Fulton, N.D.: Software rejuvenation: Analysis, module and applications. In: FTCS '95, Washington, DC, USA, IEEE CS (1995) 381
- [8] Castelli, V., Harper, R.E., Heidelberger, P., Hunter, S.W., Trivedi, K.S., Vaidyanathan, K., Zeggert, W.P.: Proactive management of software aging. IBM J. Res. Dev. **45**(2) (2001) 311–332
- [9] Hong, J., Kim, S., Cho, Y., Yeom, H., Park, T.: On the choice of checkpoint interval using memory usage profile and adaptive time series analysis. In: Proc. Pacific Rim Int. Symp. on Dependable Computing, IEEE Computer Society (2001)
- [10] Zheng, Z., Lan, Z., Gupta, R., Coghlan, S., Beckman, P.: A practical failure prediction with location and lead time for Blue Gene/P. In: Dependable Systems and Networks Workshops (DSN-W). (2010) 15–22
- [11] Cappello, F., Casanova, H., Robert, Y.: Preventive migration vs. preventive checkpointing for extreme scale supercomputers. Par. Proc. Letters **21**(2) (2011) 111–132
- [12] Ferreira, K., Stearley, J., Laros, J.H.I., Oldfield, R., Pedretti, K., Brightwell, R., Riesen, R., Bridges, P.G., Arnold, D.: Evaluating the Viability of Process Replication Reliability for Exascale Systems. In: SC'11. (2011)
- [13] Zheng, G., Ni, X., Kale, L.: A scalable double in-memory checkpoint and restart scheme towards exascale. In: Dependable Syst. and Networks Workshops. (2012)
- [14] Heath, T., Martin, R.P., Nguyen, T.D.: Improving cluster availability using workstation validation. SIGMETRICS Perf. Eval. Rev. **30**(1) (2002)
- [15] Schroeder, B., Gibson, G.A.: A large-scale study of failures in high-performance computing systems. In: Proc. of DSN. (2006) 249–258
- [16] Heien, E., Kondo, D., Gainaru, A., LaPine, D., Kramer, B., Cappello, F.: Modeling and tolerating heterogeneous failures in large parallel systems. In: SC'11. (2011)
- [17] Bougeret, M., Casanova, H., Rabie, M., Robert, Y., Vivien, F.: Checkpointing strategies for parallel jobs. In: SC'11. (2011)
- [18] Bouguerra, M.S., Gautier, T., Trystram, D., Vincent, J.M.: A flexible checkpoint/restart model in distributed systems. In: PPAM. Volume 6067 of LNCS. (2010) 206–215

-
- [19] Gainaru, A., Cappello, F., Kramer, W.: Taming of the shrew: Modeling the normal and faulty behavior of large-scale HPC systems. In: Proc. IPDPS'12. (2012)
 - [20] Fulp, E.W., Fink, G.A., Haack, J.N.: Predicting computer system failures using support vector machines. In: Proceedings of the First USENIX conference on Analysis of system logs, USENIX Association (2008)
 - [21] Li, Y., Lan, Z., Gujrati, P., Sun, X.: Fault-aware runtime strategies for high-performance computing. *IEEE TPDS* **20**(4) (2009) 460–473

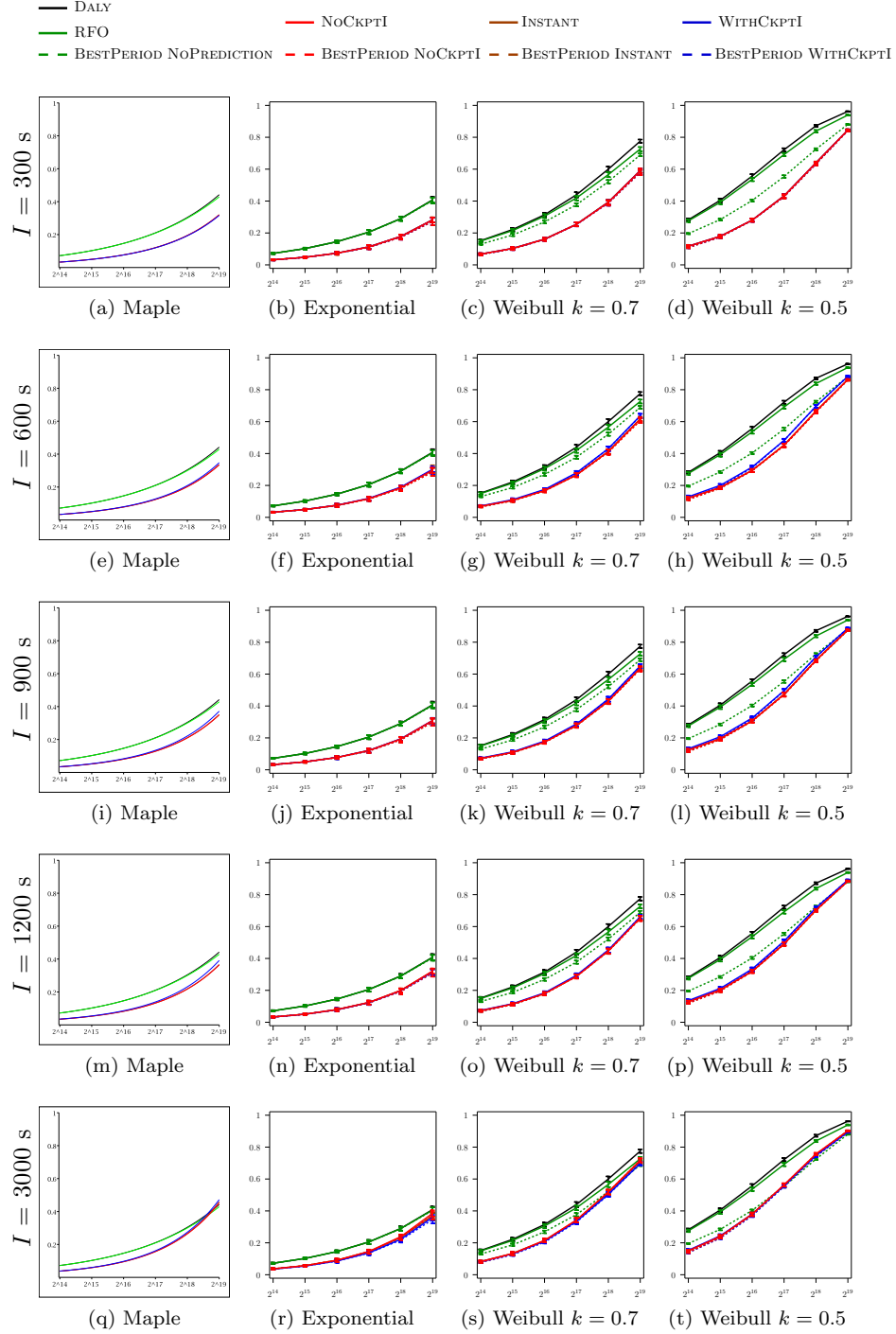


Figure 2: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

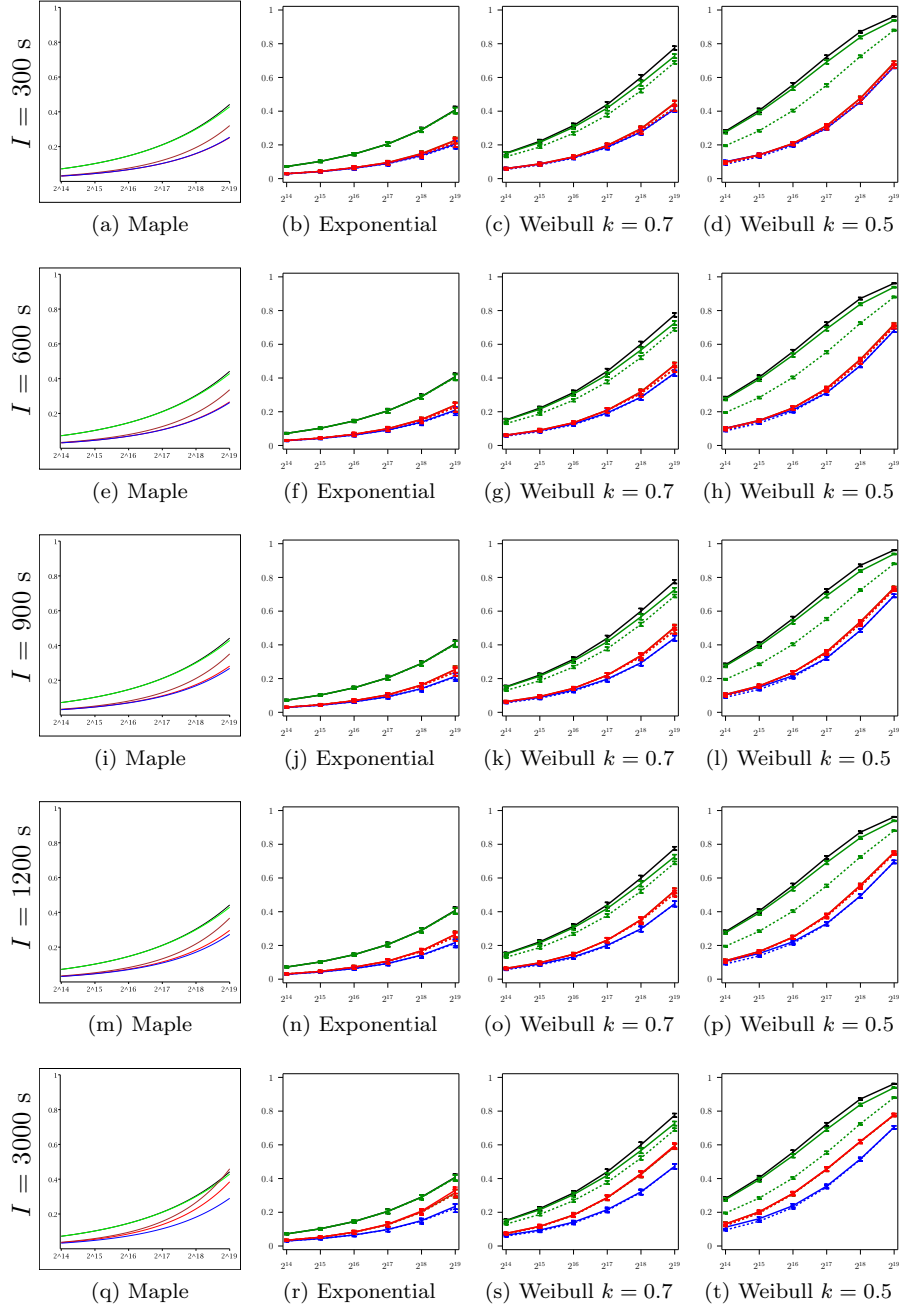


Figure 3: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = 0.1C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

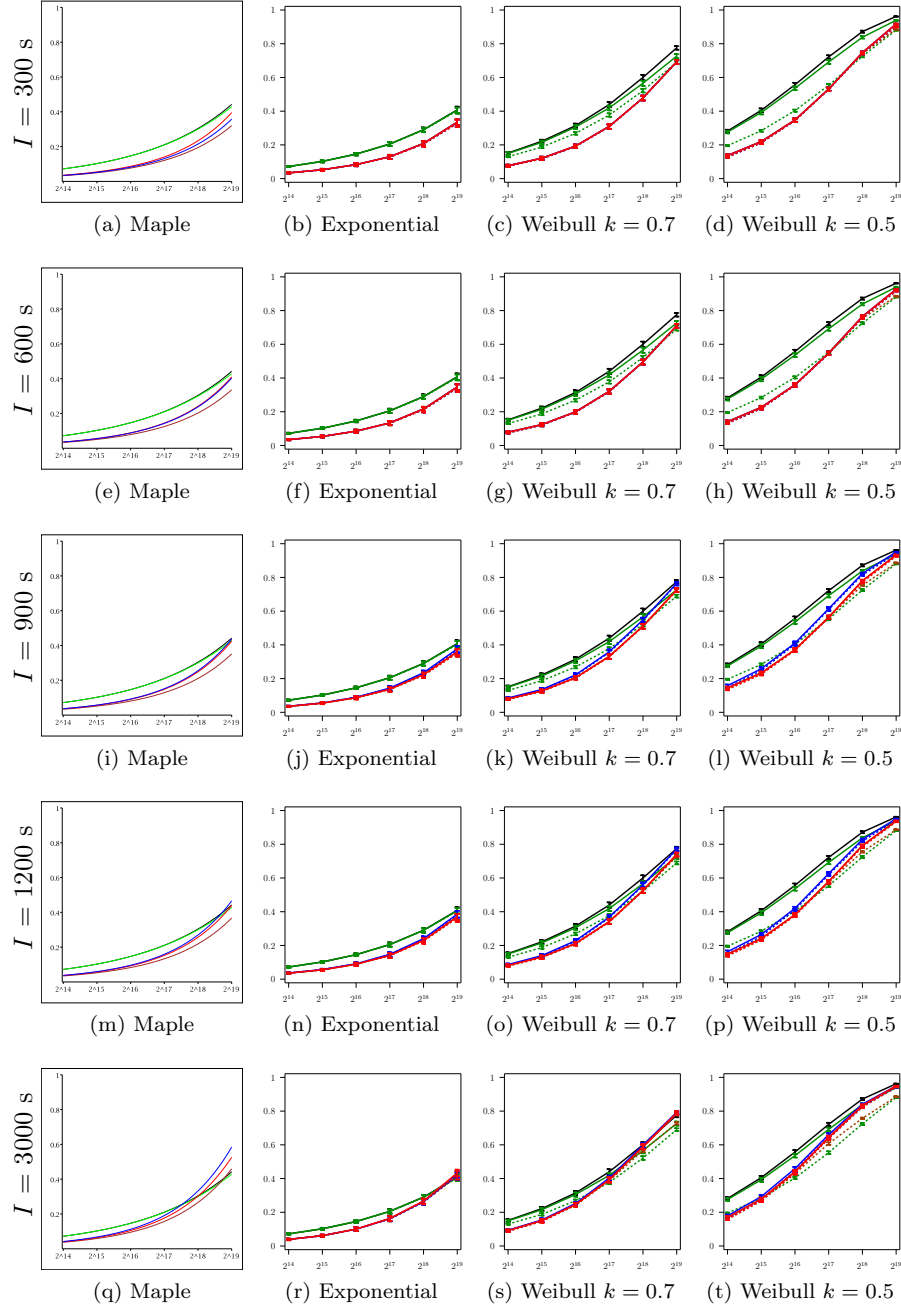


Figure 4: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = 2C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

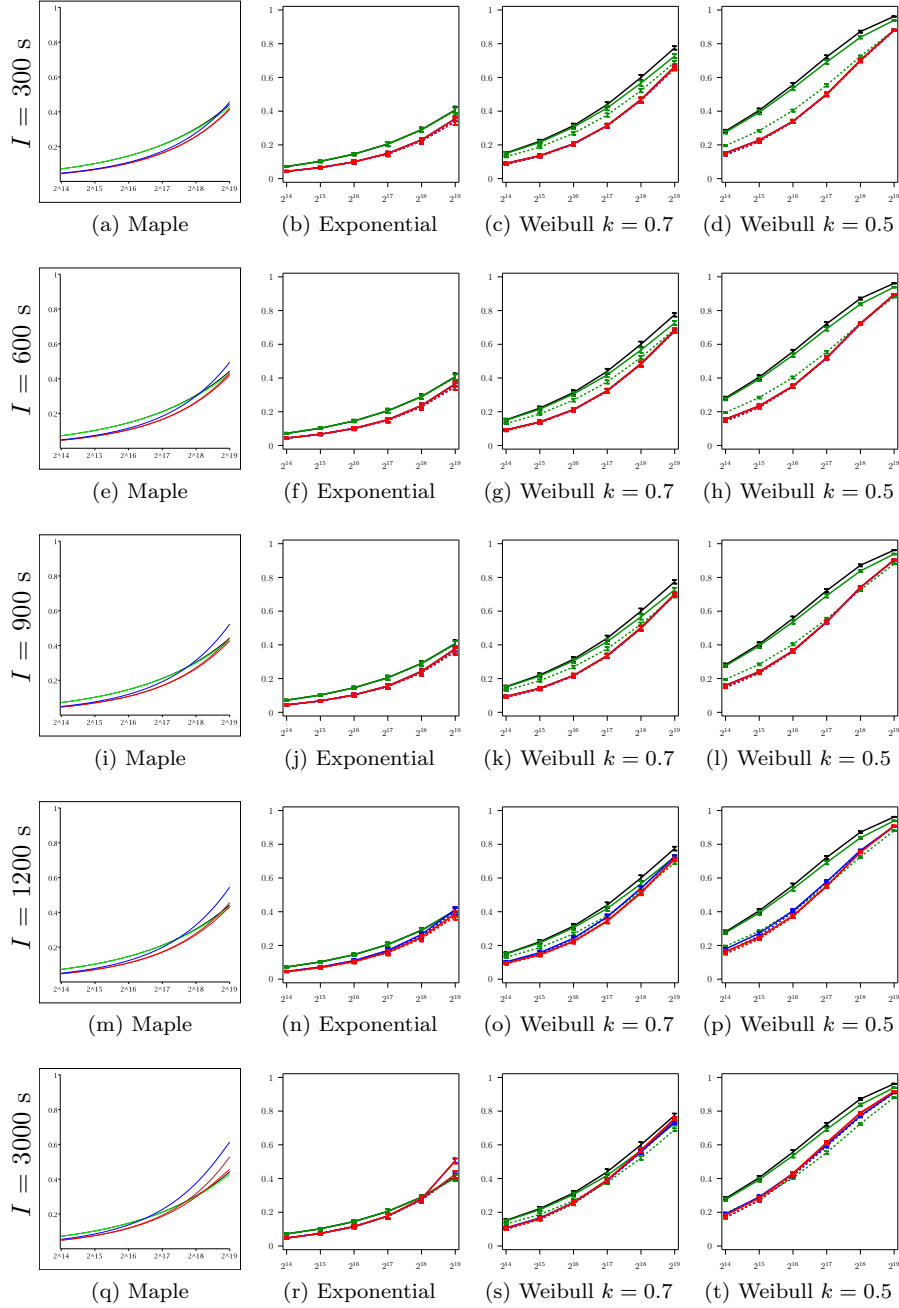


Figure 5: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

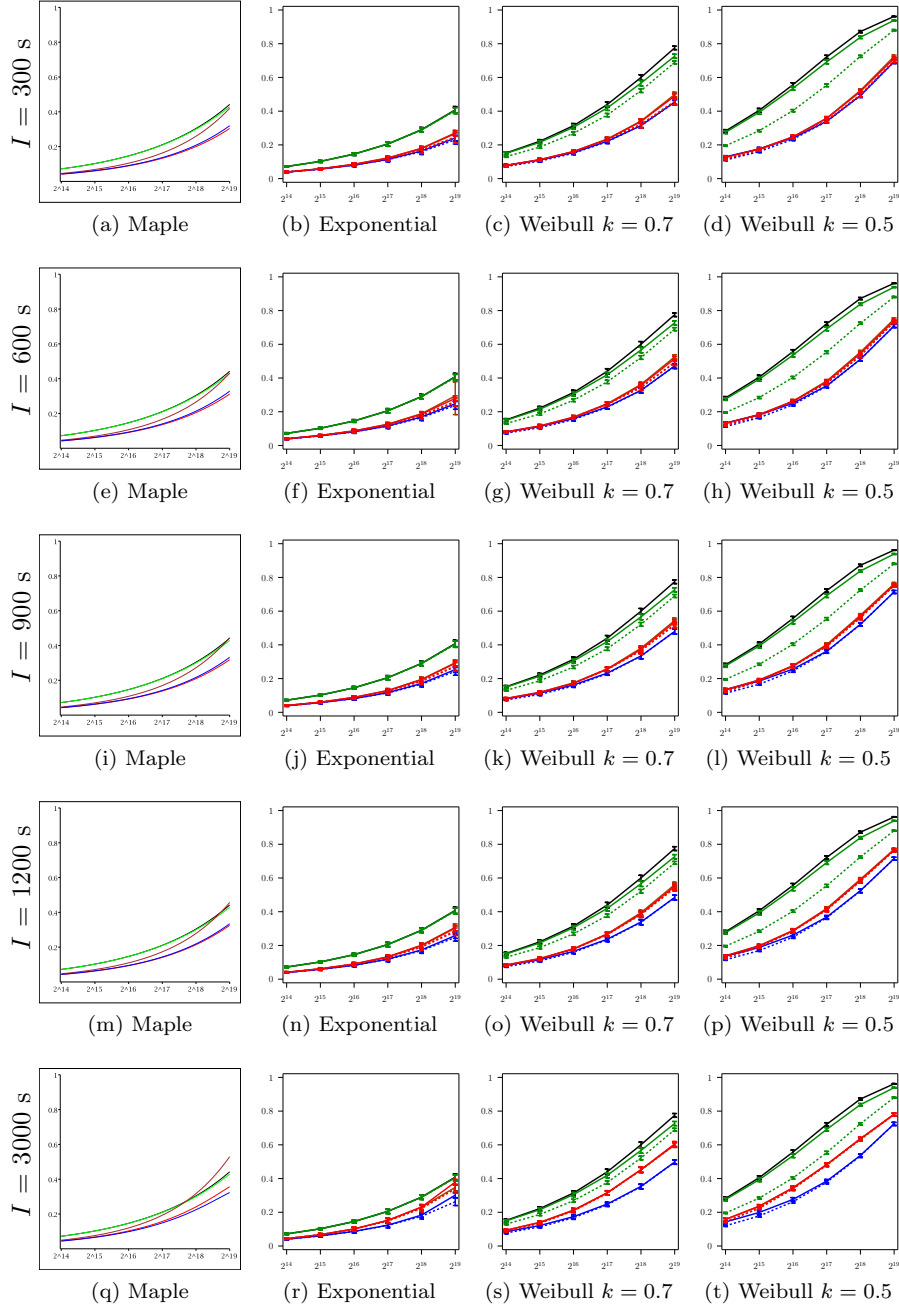


Figure 6: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = 0.1C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

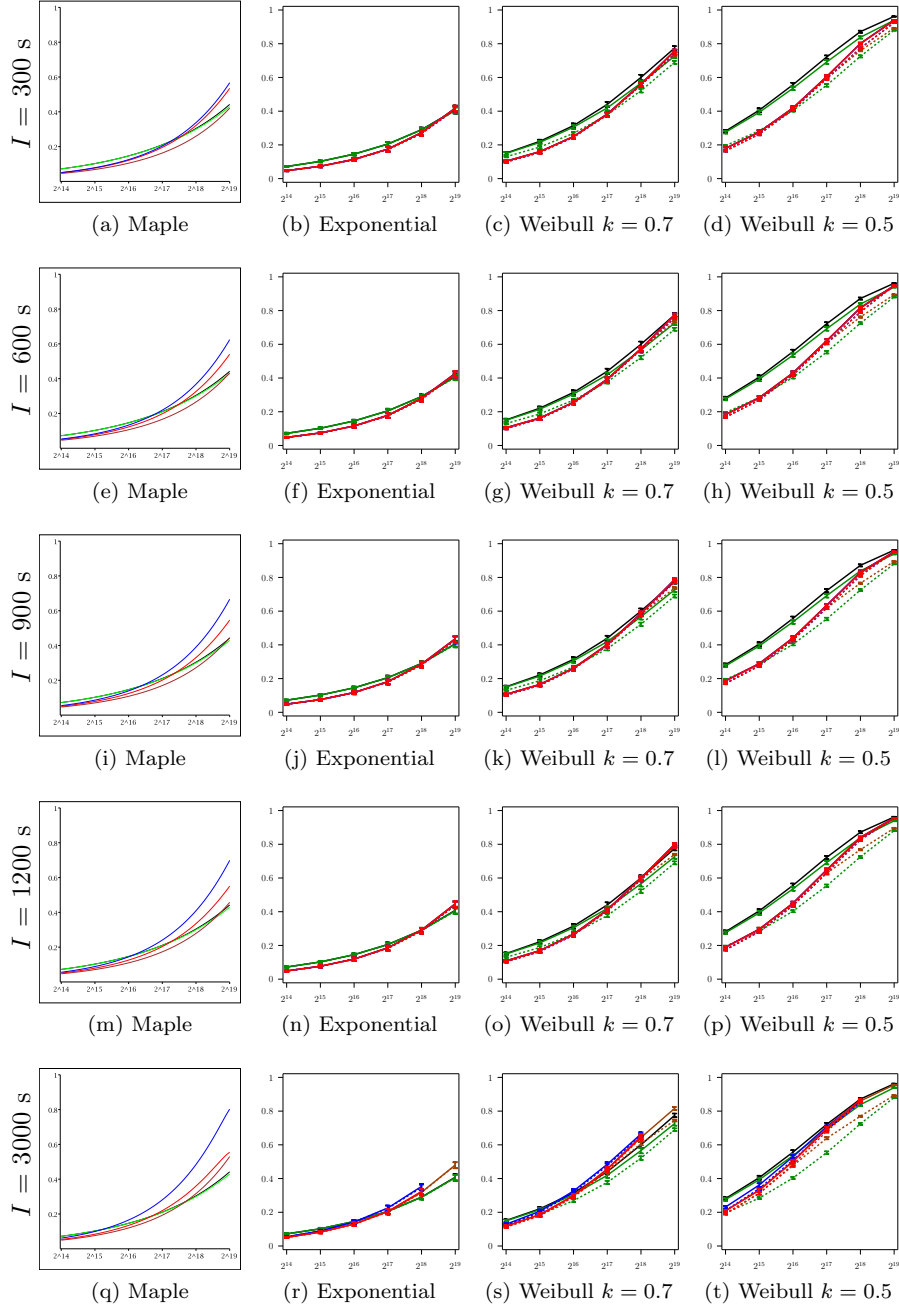


Figure 7: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = 2C$, and with a trace of false predictions parametrized by a distribution identical to the distribution of the trace of failures.

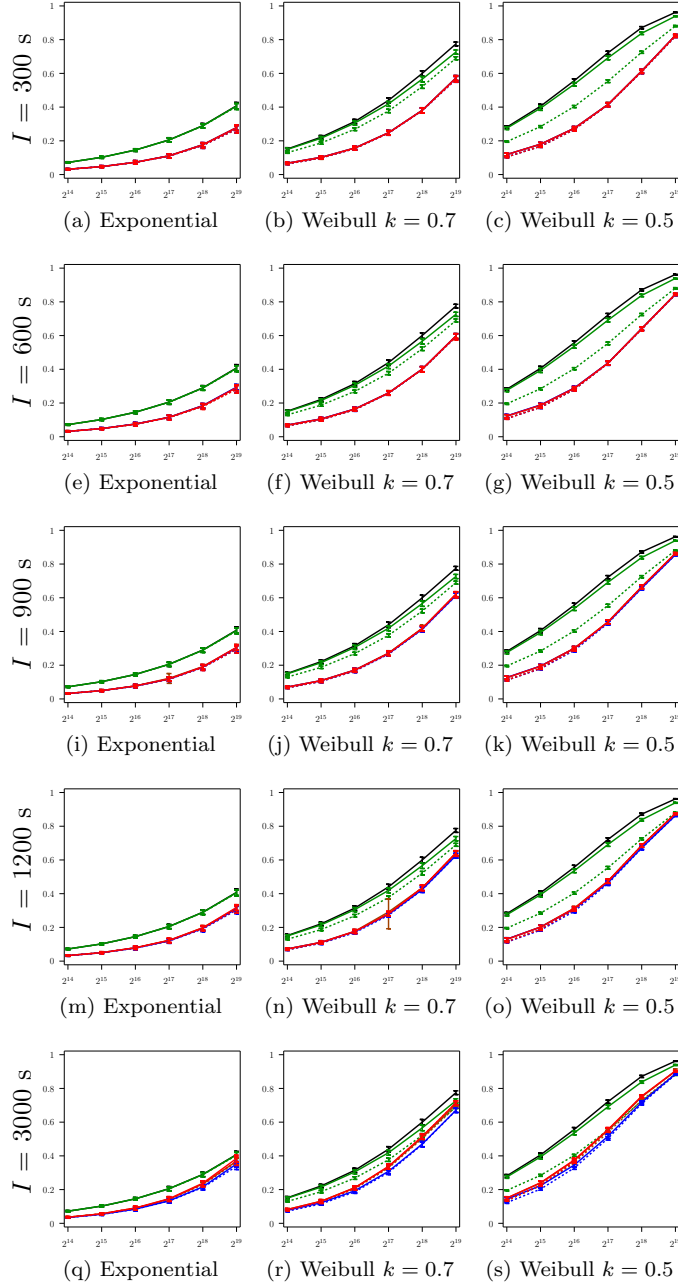


Figure 8: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a trace of false predictions parametrized by a uniform distribution.

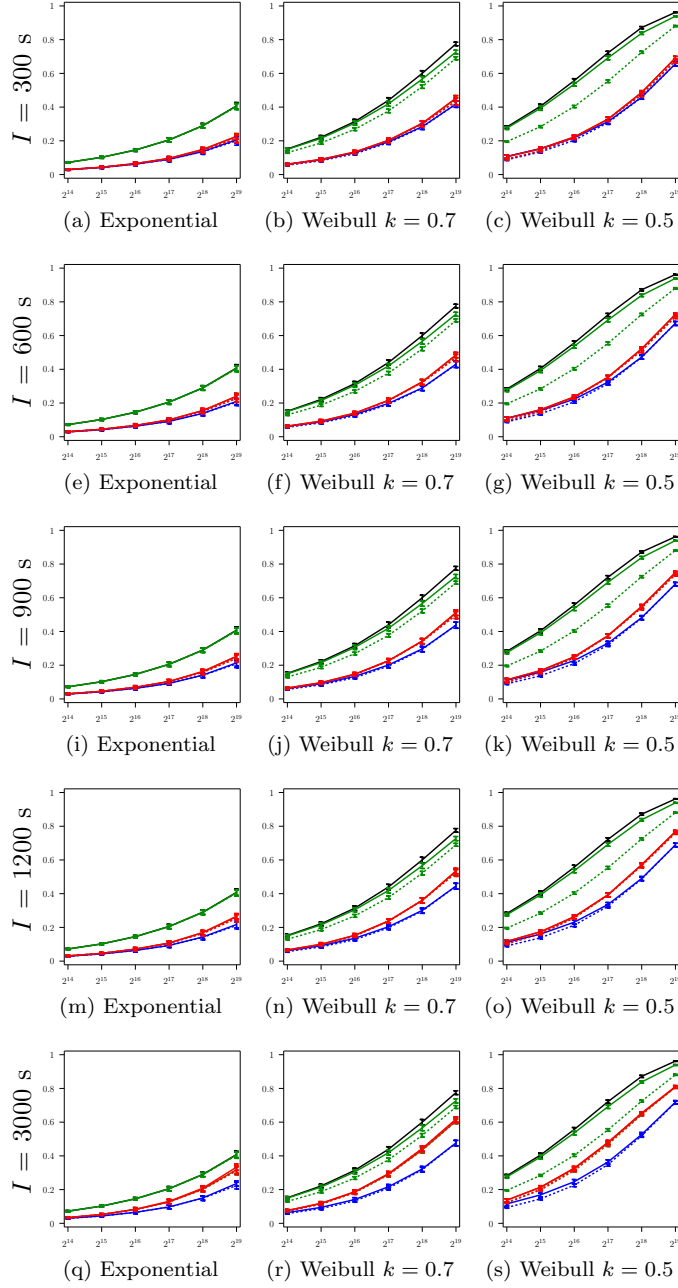


Figure 9: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = 0.1C$, and with a trace of false predictions parametrized by a uniform distribution.

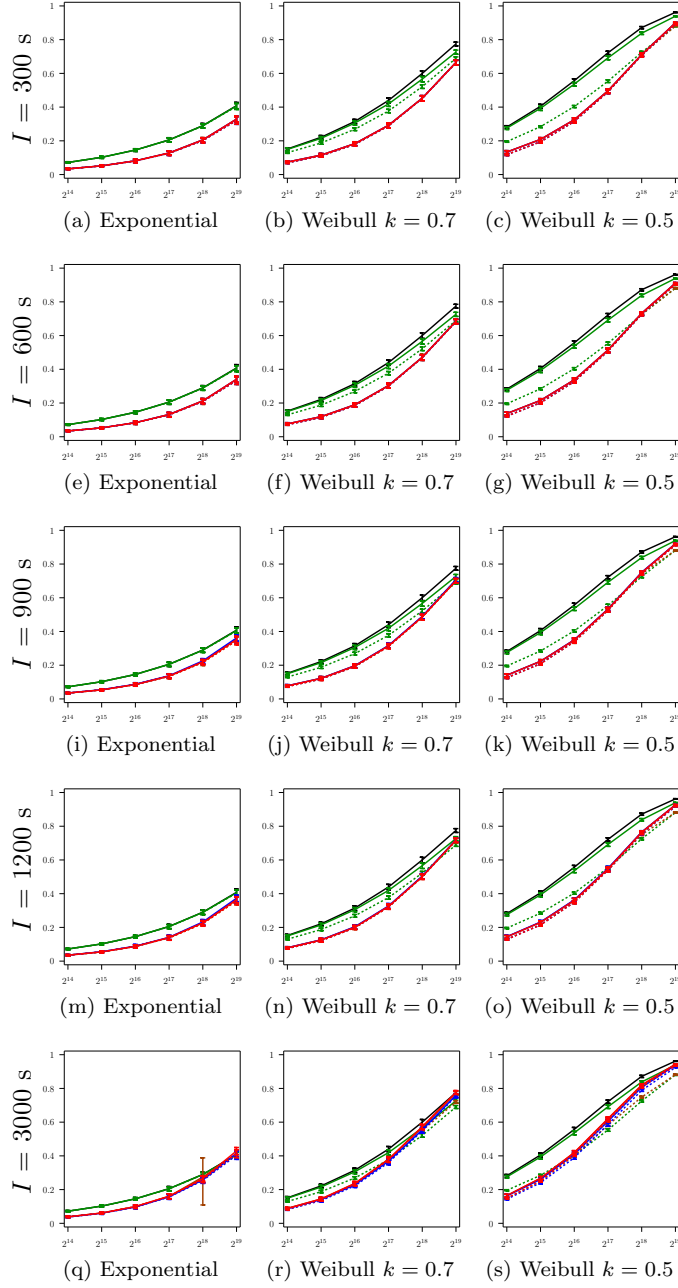


Figure 10: Waste for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = 2C$, and with a trace of false predictions parametrized by a uniform distribution.

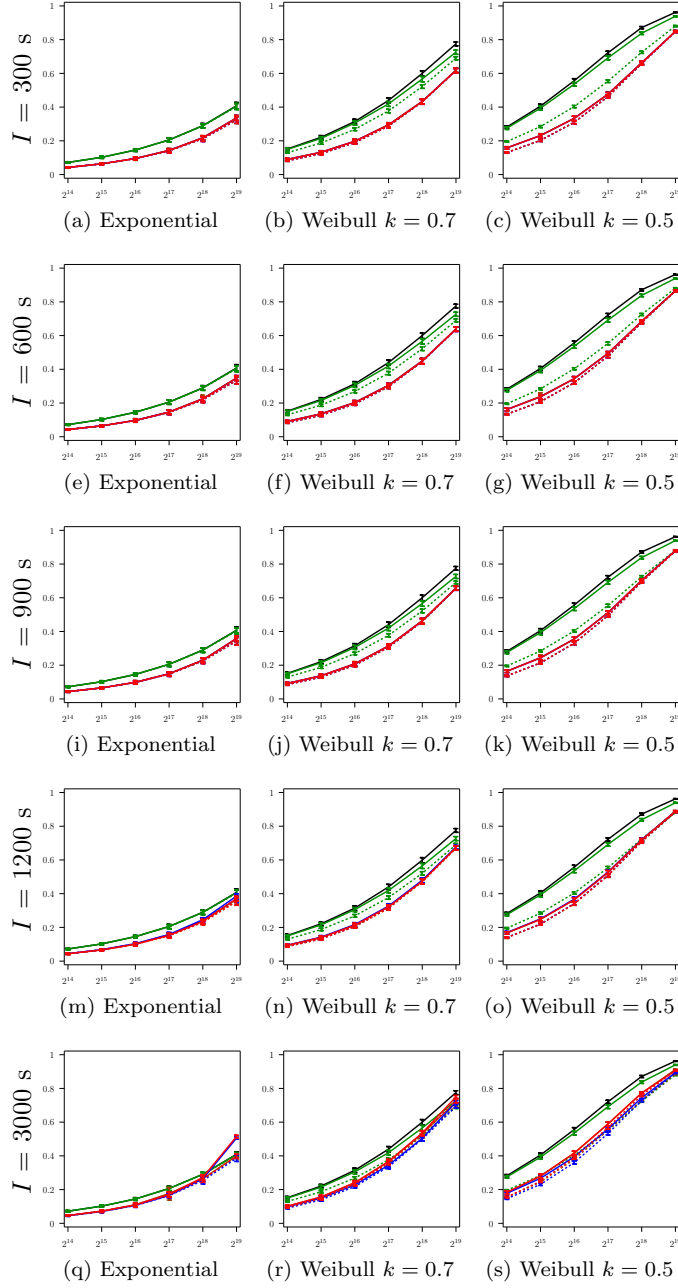


Figure 11: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a trace of false predictions parametrized by a uniform distribution.

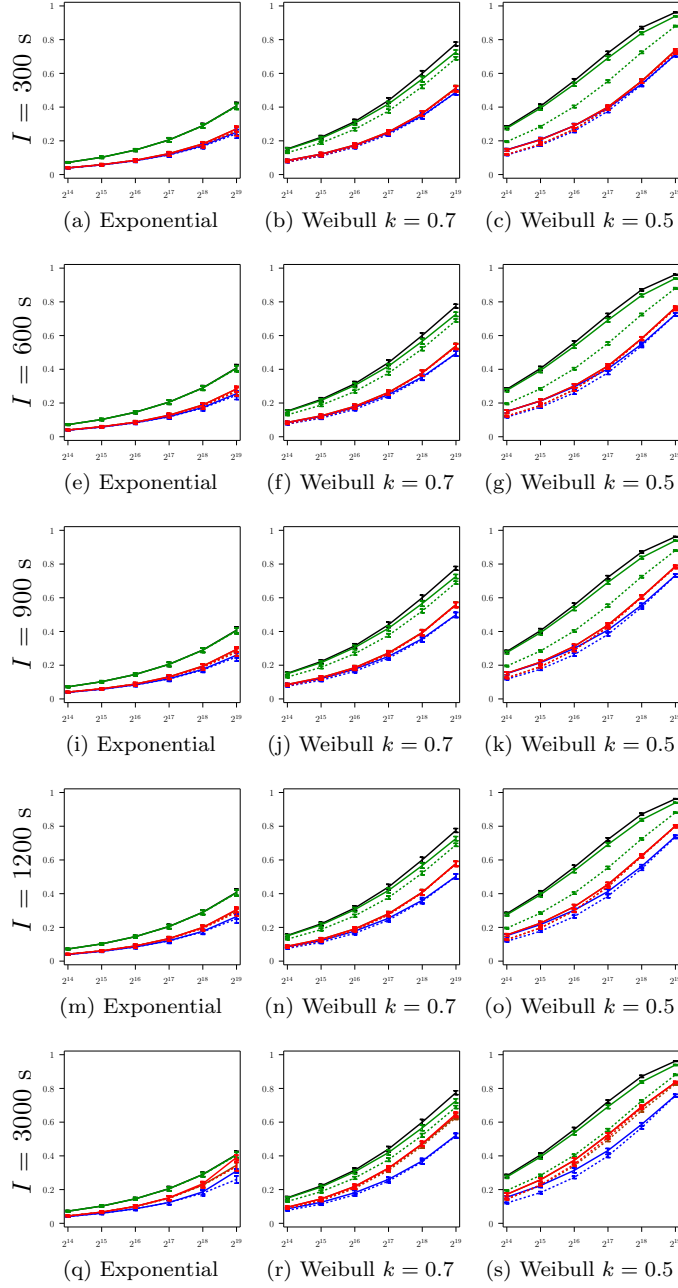


Figure 12: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = 0.1C$, and with a trace of false predictions parametrized by a uniform distribution.

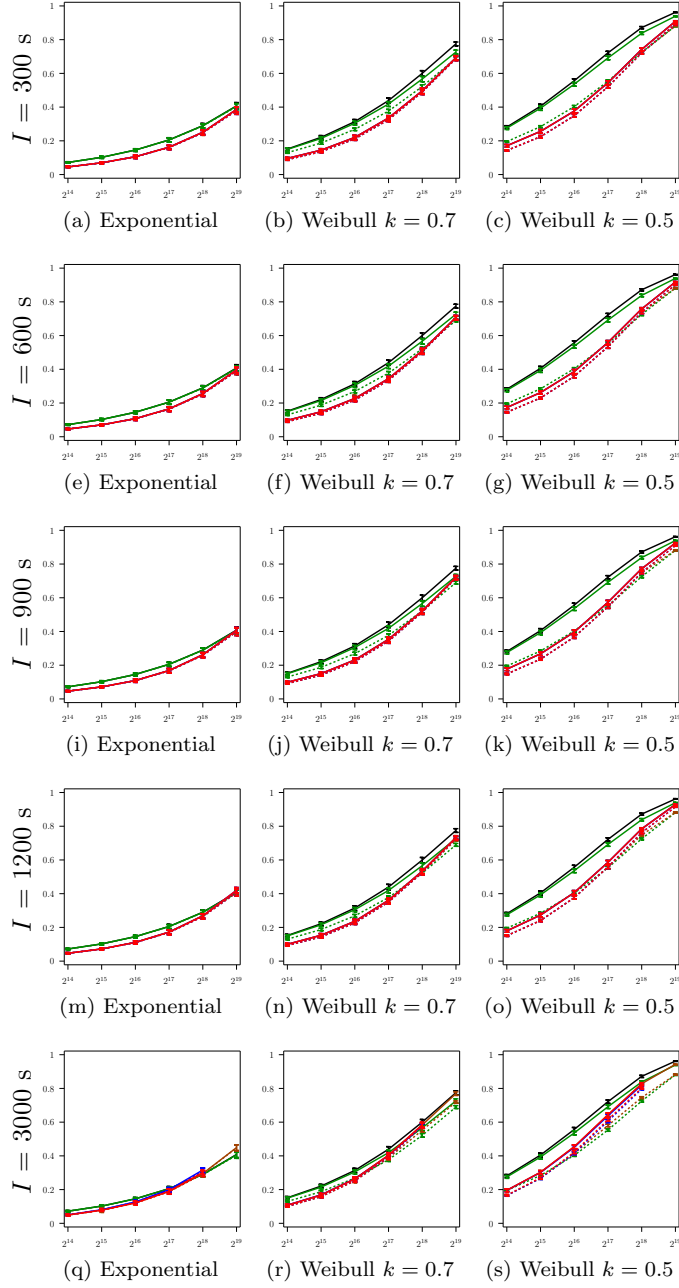


Figure 13: Waste for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = 2C$, and with a trace of false predictions parametrized by a uniform distribution.

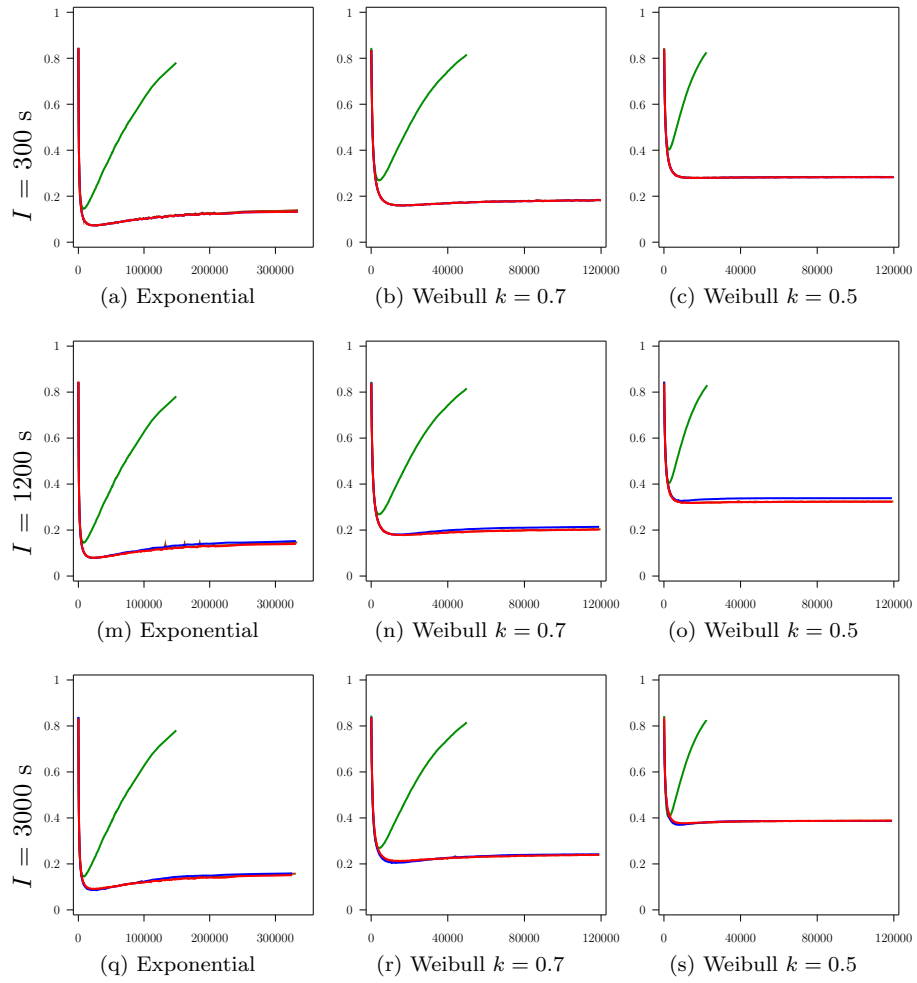


Figure 14: Waste as function of the period T_R for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a platform of 2^{16} processors.

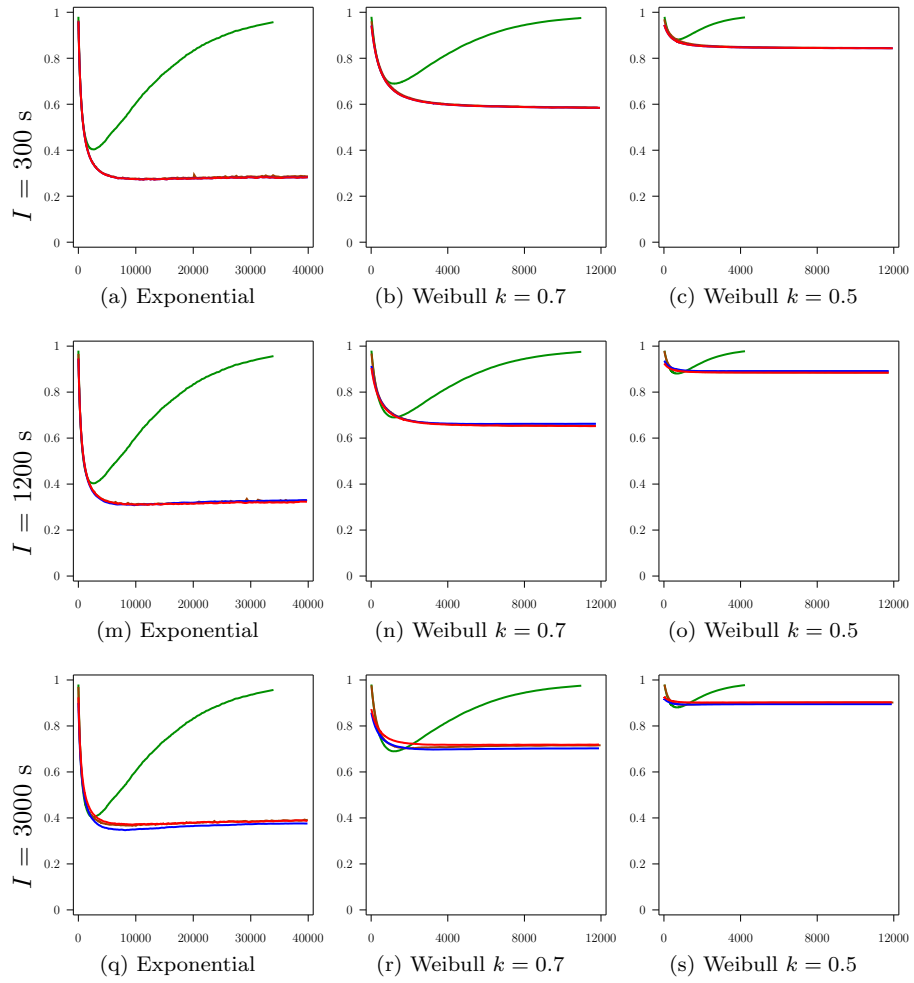


Figure 15: Waste as function of the period T_R for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a platform of 2^{19} processors.

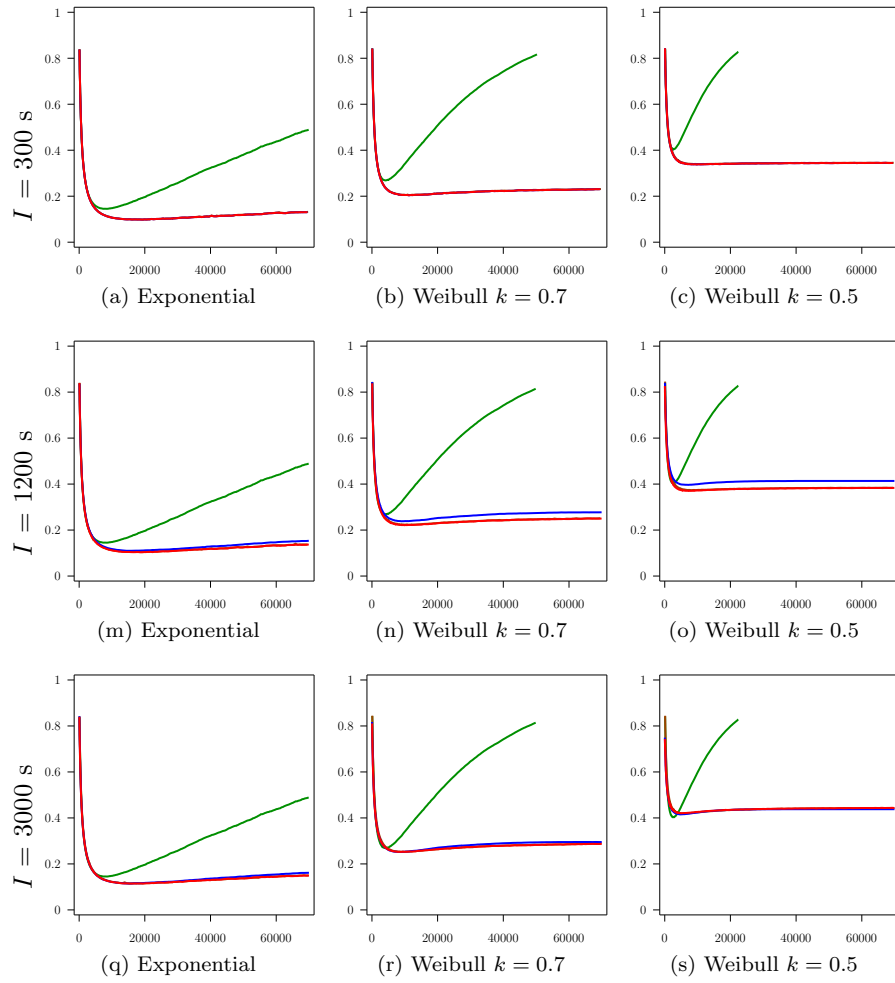


Figure 16: Waste as function of the period T_R for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a platform of 2^{16} processors.

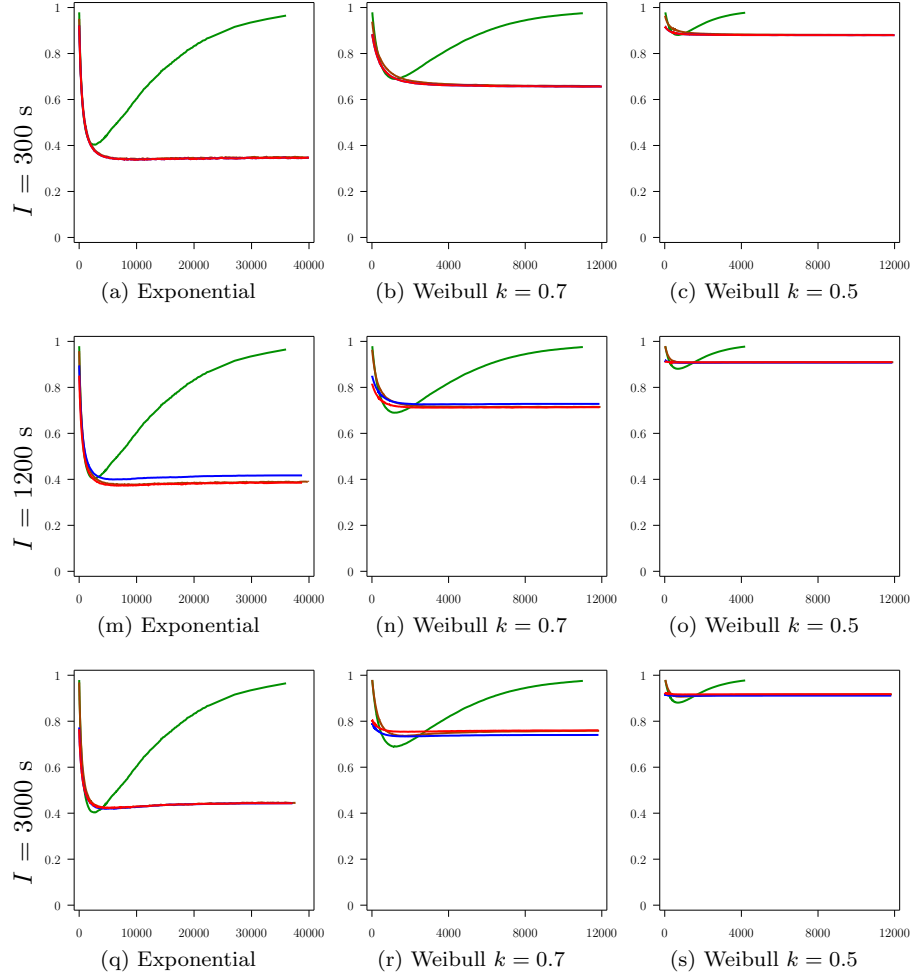


Figure 17: Waste as function of the period T_R for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a platform of 2^{19} processors.

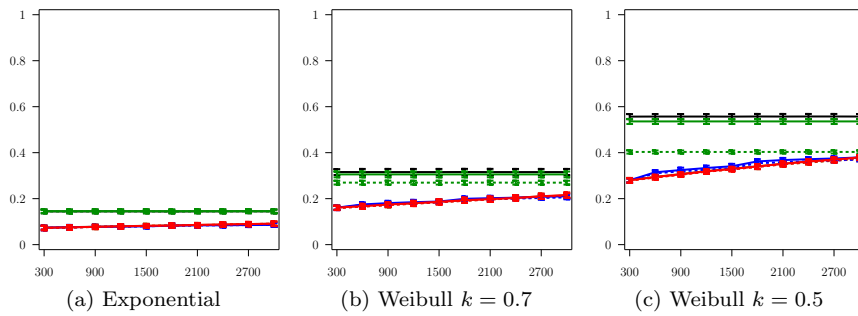


Figure 18: Waste as function of the prediction window I for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a platform of 2^{16} processors.

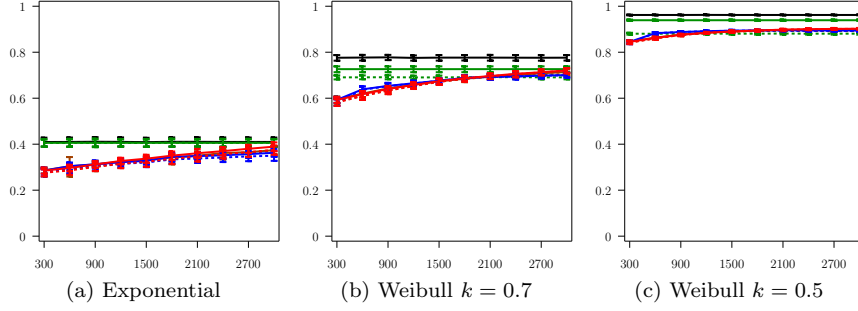


Figure 19: Waste as function of the prediction window I for the different heuristics, with $p = 0.82$, $r = 0.85$, $C_p = C$, and with a platform of 2^{19} processors.

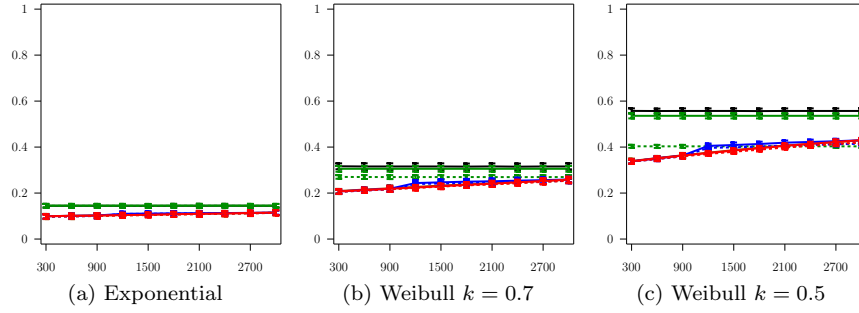


Figure 20: Waste as function of the prediction window I for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a platform of 2^{16} processors.

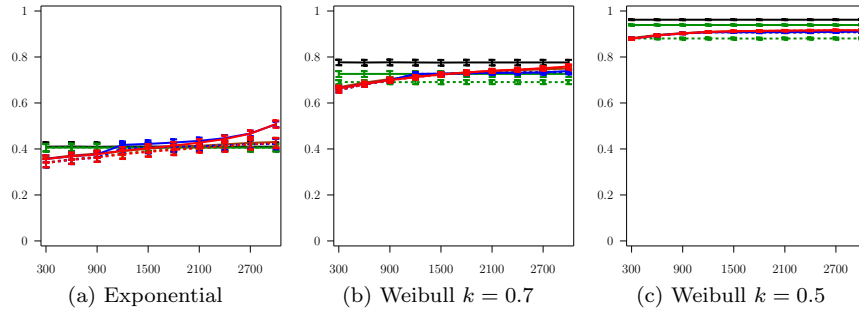


Figure 21: Waste as function of the prediction window I for the different heuristics, with $p = 0.4$, $r = 0.7$, $C_p = C$, and with a platform of 2^{19} processors.



**RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES**

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399